Fitzpatrick, Tess and Thwaites, Peter (2020)

Word Association Research and the L2 Lexicon.

Language Teaching (CUP)

Accepted for publication February 2020

STATE OF THE ART PAPER

Word Association research and the L2 lexicon

Tess Fitzpatrick

Swansea University, Wales, UK

t.fitzpatrick@swansea.ac.uk

Peter Thwaites

Keimyung University, Republic of Korea

peterthwaites@gw.kmu.ac.kr

Abstract

Since its modern inception in the late nineteenth century, research on word associations has developed into a large and diverse area of study, including work with both applied- and psycho-linguistic orientations. However, despite significant recent interest in the use of word association to investigate second language (L2) vocabulary knowledge and testing, there has until now been no systematic attempt to review the wider word association research tradition for the benefit of second language-oriented researchers and practitioners. This paper seeks to address this, drawing together applied and psycholinguistic research from the past 150 years, with a focus on research published since 2000. We evaluate the current state of L2 word association research, before identifying methodological and theoretical themes from a broader range of disciplinary approaches. Emerging from this, new paradigms are identified which have potential to catalyse a new phase of work for second-language word association scholars, and which indicate priority foci for future work.

TESS FITZPATRICK is Professor of Applied Linguistics at Swansea University, Wales, UK. Her research interests include lexical processing and retrieval, second language learning, teaching and testing approaches, and the extension of SLA research methods to other contexts, and she has published extensively in these areas.

PETER THWAITES is an Assistant Professor of English Education at Keimyung University, Republic of Korea. His research focuses on the psycholinguistic underpinnings of associative lexical knowledge, in both first and second languages. He has also published in the area of second language writing pedagogy.

Corresponding author: Peter Thwaites, Keimyung University, Yeung-am Gwan219, 1095 Dalgubeol-daeru, Dalseo-gu, Daegu 42601, Republic of Korea. peterthwaites@gw.kmu.ac.kr

Contents

1. Introduction

Word association research is deceptively simple. In its most basic form, it involves presenting participants with cue words, and asking them to respond with the first word that comes to mind. This type of data is quite easy to collect, and this, together with the technique's long history as a method for researching L1 lexical knowledge, has made it an attractive approach for researchers interested in understanding what second language learners know about words. Consequently, the word association (WA) technique has been used as a basis for exploration of the content and structure of the L2 lexicon, as a method for understanding the bilingual brain, and as a technique for assessing vocabulary knowledge, amongst other research aims.

However, from the apparent simplicity of the WA technique has emerged a wealth of potential analytic approaches and, partly as a consequence, L2 WA research has not identified consistent behaviour patterns. Nonetheless, notions of 'promise' and 'hope' persist (Schmitt 1998a; Wolter 2002; Zareva & Wolter 2012), and researchers continue to refine WA protocols and analyses in pursuit of meaningful patterns. In this State of the Art paper we tease out evidence of that sense of unrealised potential, and examine reasons behind it. We draw together and juxtapose themes and findings from subsets of research, and use this to suggest new ways of approaching WA research, and to predict future developments.

Word association tasks have been used since the late 19th century to investigate conceptual and lexical connections in the domains of psychology and psychiatry and, increasingly, in connection with applied linguistics (see Section 2). The latter context hinges on the fact that a WA task typically requires a participant to respond to a lexical cue with the first word that comes to mind, and the resulting data therefore carries potential information about lexical availability, organisation and retrieval. WA research in the 1960s revealed shifts in association behaviour during L1 development, and this generated interest in the capacity of

WA to measure second language proficiency. A significant body of applied linguistics research in the last fifty years has attempted to develop analytic approaches to realise and exploit this (see the annotated bibliography in Meara 2009: 101–127). Meanwhile, studies in experimental psycholinguistics have used WA tasks to investigate the roles of automaticity and conceptual memory in lexical retrieval behaviour. Associative retrieval, both as an aspect of lexical knowledge and as a means of investigating automaticity, has direct relevance to our understanding of language acquisition and processing, and by extension to language teaching and assessment issues. However, until the current paper there has been no comprehensive account of the relationship between research in these two domains (applied- and psycho-linguistic), nor of ways in which findings might interact in a mutually informative manner.

This article maps out, and attempts to redress, the somewhat fragmentary research landscape of WA, and its publication is timely for two reasons: First, chronological database analyses of publications related to WA and linguistics/language demonstrate periodic surges of interest in this topic[i]; they are evident in the first two decades of the 20th century, in the 1960s and 70s, and in the most recent two decades, marking this as a judicious time to take stock. Second, current and emerging technology i) facilitates WA research in that the big data network perspectives called for by Deese (1962b) can now be realised, and have been, for example in the work of De Deyne and colleagues (De Deyne et al. 2018, and see Section 5); and ii) prioritises WA research in that linguistic input to the natural language processing (NLP) mission demands nuances of semantics and usage that go beyond conventional NLP resources, but that might be mined from careful analysis of WA data. Applying technological developments to WA data can have positive consequences for second language learners and users: it can increase the sophistication of language training tools, enhance machine translation, and potentially create mental language maps to augment the corpus information that transformed language learning and teaching in the late 20th century.

Our central focus here, L2 WA research, is with a couple of exceptions a relatively new phenomenon. In order to understand and assess the propositions on which it is based we begin, below, by examining landmark contributions to the longer WA research tradition. After this orientation, we scrutinize the WA research that set out to investigate second language acquisition and use, and which falls broadly into two strands: a focus on knowledge of individual lexical items, and a focus on storage and retrieval routes in the lexicon. Methods of investigation vary considerably and are rife with assumption, oversight, and operational challenges. We identify three key notions in WA research that can help in addressing these shortcomings: the network metaphor, the influence of specific lexical variables, and models of psycholinguistic processing. This equips us to sift and make sense of the most robust and salient findings in recent and current L2 WA research, and to predict and propose future L2 WA research directions.

2. The word association research tradition

A surprisingly large proportion of WA research publications begin with a historical orientation, and a consequence of this has been a kind of snowball citation effect, whereby some of the early literature is referred to often, and some hardly at all. This selective focus has promoted, in the applied linguistics context, an overly simplistic sense of a linear research pathway for WA, from psychoanalytic research in the early twentieth century, to broader psychological work including child cognition around the 1960s, to (applied) linguistics research including L2 investigations more recently. This view obscures early language-oriented work, and can discourage researchers from looking across chronological and disciplinary boundaries to inform their work. In order to avoid this pitfall in the current paper, we conducted systematic searches in the 'Scopus' and 'Linguistics and Language Behaviour Abstracts' databases using keywords *word association*, linguistic** and

*language**. In this section we consider early intimations of themes that are prevalent in recent and current WA work, and that help us to contextualise and evaluate the degree to which recent findings have progressed the field.

In an 1879 edition of *Brain*, Galton describes an experiment, with himself as sole participant, designed to capture and measure 'associated ideas'. It entailed the generation of 75 random words, and of four subsequent sets of two associations for each of these words, produced in a timed condition. He categorised response behaviours into i) 'instantaneous… just as a machine might act'; ii) 'sense-imagery… visual imagery', and iii) 'histrionic' with a 'nascent sense of some muscular action' (1879: 159); these resonate, respectively, with the i) collocative / SYNTAGMATIC; ii) meaning-based / PARADIGMATIC / conceptual; and iii) EMBODIED COGNITION distinctions seen in much more recent work (see Sections 4 and 5.3). Galton classified his cue words too. His categories are not labelled, but examples indicate that they are concrete nouns, emotion words, and abstract words. In a 'comparison between the quality of the words [cues] and that of the ideas [responses] in immediate association with them' (1879: 160) he suggests that the former influences the latter. In this, together with his hints that association patterns change with age, that RESPONSE TIME is informative, and that associations are not straightforwardly biddable (the cue 'abasement' prompted associations to 'a basement'), Galton presages questions still confounding WA research today.

Galton was a polymath, without affiliation to any particular discipline, and presented this 1879 work as 'new'. However, within thirty years the WA method was seen as 'in vogue in psychology, [and] so familiar to [the audience at Clark University] that there is no need to speak of it.' (Jung 1910: 219) . It is in this psychology tradition that the early WA work most familiar to today's linguists emerged. As well as the unquestionable cultural legacy of this work, its methodological impact is persistent, and now transcends disciplinary boundaries: the cue words from Kent and Rosanoff's (1910) *Study of Association in Insanity* are still

widely used – Google Scholar reports well over 300 citations of their work since 2000, with around half of these in linguistics-oriented outputs (accessed 31 October 2019). What is less widely acknowledged is that the cues used by Jung and by Kent and Rosanoff were translations of German word lists (e.g. Sommer 1901), selected originally for experimental work in that language. Any apparent lack of concern in those scholars about the provenance of their cues stands in contrast to the approach of Esper who, also over a century ago, set out explicitly to replicate in English work conducted by Thumb & Marbe (1901) in German, in order to assess whether association patterns systematically differed between the two languages (Esper 1918). Esper's findings are unarguably framed in linguistic terms, and anticipate those in much more recent studies: he found that i) in both languages reaction times are faster for more frequently produced responses; ii) educated adults produce similar associations to those of uneducated adults and children, but produce them faster; iii) the word class of the cue word tends to be reflected in the response. He also notes that Thumb & Marbe considered a response produced 'spontaneously', 'without any intervening mental process' to be 'linguistically effective', meaning that it has the capacity to create analogous forms or meanings; this distinction between truly spontaneous and other WA responses was not systematically pursued again in WA research until very recently (see Section 5.3), though Woodworth revisited it in his information-packed 1938 state-of-the-art chapter "Association" in *Experimental Psychology*. In contrast to Kent and Rosanoff's 300+, Esper has just nine post-2000 citations (Google Scholar, accessed 31 October 2019).

The next significant boost to language-focused WA research came in 1952 when the US Social Science Research Council set up a *Committee on Linguistics and Psychology* to investigate language behaviour. This generated, among other things, a renewed interest in WA methods, and by the 1960s these had largely crystallised around two fairly distinct paradigms, which account for most subsequent WA research and which are addressed in the

next two sections of this paper. In the first paradigm (see Section 3), collections of NORMS LISTS are created from individuals' WA responses, and are used to investigate and compare the associations of specific linguistic communities (see Postman & Keppel 1970). These data are typically used to investigate three kinds of variable: features of the individual respondents (e.g. gender, age, class, first or second language user), lexical features of cues and/or responses (e.g. frequency, homonymy), and cross-language comparability, a concern that is somewhat absent from recent research, but which is exemplified in Rosenzweig's (1961) investigation of translation equivalence of primary responses in German, French, Italian and English, and which Esper had noted (see above).

The second paradigm (see Section 4) applies Saussure's notions of paradigmatic and syntagmatic relationships to WA data. Researchers of child language development detected a shift from syntagmatic to paradigmatic responses at around 6-8 years old (e.g. Brown & Berko 1960; Ervin 1961; Entwisle, Forsyth, & Muuss 1964; Palermo 1971). This was initially taken to indicate a dynamic restructuring of the lexicon, but other interpretations were subsequently proposed, accommodating such notions as cognitive processing (Francis' 'thoughtful operations of comparison and inclusion'; 1972: 957), conceptual reorganisation, and task capacity (K. Nelson 1977). Nelson's note that 'the syntagmatic-paradigmatic shift, once seemingly so clear and reliable, has tended to dissolve or resolve into unanticipated, complicating elements' (1977: 114) highlights the contrast between the simplicity of the WA methodology and the complexity of its interpretation. One fundamental challenge is to distinguish between cue word characteristics and respondent characteristics, in terms of their contribution to determining a WA response. This had already been noted by Deese (1962); for him, the only way to address this tension was by focusing on the observation that 'associations exist in well organised …. networks' (1962: 163) and by tracing association distributions through large WA data sets. This idea is explored further in Section 5.

In 1980, Meara's survey article on vocabulary acquisition positioned WA research explicitly within the domain of language learning by relating WA studies between 1956 and 1980 to second language acquisition. Here, and in more detail in his monograph *Connected Words* (2009), he notes that early attempts to relate WA response behaviour to L2 proficiency used a range of indicators, including i) the number of responses produced/number of blanks (e.g. Riegel, Ramsey, & Riegel 1967), and ii) the number of L1 responses to L2 cues (Rüke-Dravina 1971). In general, though, he finds that investigations were dominated by the two approaches – norms lists and response categorisation – that we have identified above, expressed in two broad hypotheses

- as proficiency increases, the items produced in response increasingly mirror those produced by L1 speakers[ii] of the target language, and
- as proficiency increases, the types of response given will systematically change.

It is perhaps surprising that in this paper, positioned as a *State of the Art* piece, we have begun by giving attention to work produced so long ago. Our justification is twofold: First, the notion of State of the Art may suggest incremental but inevitable progress towards deeper or more accurate understanding, whereas in fact WA research is plagued by blind alleys and Gordian knots, with initially promising lines of enquiry failing to produce consistent findings, and layers of variables masking connections and patterns. In this context the early work can provide valuable opportunities for (re)orientation. Second, the themes identified above – norms list approaches, categorisation approaches, networks, lexical variables, response time significance - are enduring elements of L2 WA research. However, they are often considered in a fragmentary fashion, and keeping early studies within our frame of reference ensures that recurrence of phenomena or propositions will not be missed.

3. Using word association to investigate knowledge of vocabulary items

The body of WA research most transparently related to language learning and teaching is based on the idea that association is an aspect of word knowledge. Nation's word knowledge framework (Nation 1990; 2001) and Richards' lexical competence taxonomy (1976: 79–81) capture this notion explicitly: Nation lists ASSOCIATIONS as one of nine components of 'What is involved in knowing a word?', and one of Richards' eight 'assumptions' about lexical competence is:'knowledge of the network of associations between [a] word and other words'. Richards makes specific reference to the informative data produced in WA tasks. Both he and Nation focus mostly on semantic associations, but there is an acknowledgement that different kinds of association are likely to be salient at different stages of learning. Indeed, WA data has been used to capture information about definitional, collocational, orthographic, morphological knowledge and so on. Fitzpatrick, for example, uses WA as an elicitation tool to track acquisition of these aspects of knowledge and to gain 'an insight into the micro-development of these lexical features' (2012: 93). Connotative information often revealed in WA responses can denote traces of nascent or attriting word knowledge and may exist before or after definitive form-meaning links are active. Zareva's exploration of 'frontier words' (2012) relates to this, as does Meara's suggestion that Richard's association 'assumption' differs from the other seven in that it is psycholinguistic rather than descriptive in nature, and focuses on the integration of a word into an existing lexicon (Meara 1996a). We return to these ideas later in this paper, where we examine approaches that use association categories to detect patterns of lexical organization (Section 4), and studies using WA response times to evaluate speed/automaticity of access (Section 5.3).

Most typically, measurement of association knowledge is operationalized by comparing L2 learner responses with 'STEREOTYPICAL' responses (usually from L1 NORMING data; responses from groups of L1 participants, ranked according to most frequently-given

response to create norms lists). The comparison of learner and L1 speaker data resonates with Meara's call for methods of learning that produce L2 users with WA responses similar to those of first language speakers (1978: 211). He argues that such methods will by definition nurture word knowledge that goes beyond the matching of translation equivalents, and incorporates multiple nuances and uses. There has been a decline, since the mid 1990s, in the number of studies underpinned by the assumption that proficiency maps straightforwardly onto production of L1-like responses. However, the notion of RESPONSE STEREOTYPY as a gauge of competence remains, often alongside measures of the NUMBER OF RESPONSES produced to a cue item. A related research strand, using the WORD ASSOCIATES FORMAT, examines capacity to recognise, rather than produce, a word's associates from a closed set of items. In 3.2 and 3.3 below, these proposed measures of vocabulary knowledge are explored futher, but we begin with an overview of claims from the research literature about the ways in which WA behaviour relates to word knowledge and other aspects of language proficiency.

3.1 Main claims about the relationships between WA, word knowledge, and general proficiency

Despite a frustrating lack of reliably consistent findings in WA research, a number of claims about the way WA relates to proficiency have been generated. The strength of these claims varies, and some findings are contradictory (marked ! below). Later in this paper we examine in more detail the methodological approaches and analyses underlying these claims. First, though, we note some indicative claims, with examples of studies supporting them:

- Number of responses produced –

- As learners become more proficient there is an increase in the number of responses they produce to cues in a multiple-response WA task (Lambert 1956; Randall 1980; Zareva, Schwanenflugel & Nikolova 2005)
- As learners become more proficient, they will increasingly resemble L1 speakers in the number of responses they provide to specific cues (on the basis that some cues are more 'provocative' – generate more responses - than others) (Lambert 1956)

! Vocabulary measures correlate more strongly with the number of responses learners produce, than with stereotypy scores (Zareva 2005)

! Proficiency measures (cloze test and TOEIC scores) correlate *less* strongly with the number of responses learners produce, than with stereotypy scores (Munby 2018)

- Similarity to L1 norms (also referred to as stereotypy, native-like commonality)
  - As learners become more proficient, most produce slightly more responses that are the same as those of L1 speakers (Randall 1980; Schmitt 1998b, Fitzpatrick 2012)

  ! Similarity to L1 norms correlates moderately with cloze test scores (Kruse, Pankhurst, Sharwood Smith 1987; Wolter 2002), TOEIC scores (Munby 2018) and vocabulary knowledge scores (Zareva 2005)

  ! There is no significant correlation between similarity to L1 norms and grammar monitoring test scores (Kruse et al. 1987)

  ! Similarity to L1 norms could distinguish between L1 speakers and learners, but could not detect differences in proficiency level (Zareva et al. 2005; Zareva & Wolter 2012)

- WEIGHTED and UNWEIGHTED scoring systems

- Scoring systems that account for the position of a response on L1 norms lists (weighted scoring) are not significantly more sensitive to learner proficiency than those that don't (unweighted scoring) (Kruse et al. 1987; Wolter 2002)

- Reliability of WA measures
  - There is a moderately strong test-retest correlation for number of responses produced and more modest correlations for stereotypy (Kruse et al. 1987)
- Receptive knowledge of WA (as determined by the word associates format (WAF) approach, whereby participants identify items that are associates of the cue/target)
  - Rasch analysis indicates that the WAF test is reliable (high person separation reliability) (Read 1998)
  - There is a strong correlation between WAF scores and reading and vocabulary test scores (Read 1993, 1998; Qian 2002; Qian & Schedl 2004)
    - ! High and low WAF scores correspond with performance in a vocabulary-focused interview (Schmitt, Ng and Garras 2011)
    - ! WAF performance can be contradictory, with learners selecting a combination of correct and incorrect responses to some items (Schmitt et al. 2011)

3.2 Word association 'stereotypy' research in L2 – methodological approaches

A significant proportion of L2 WA research focuses on comparing learners' responses with those of expert users of the language – in most cases these being L1 speakers. In this section we examine the methods used in studies that attempt to assess learner proficiency by examining the stereotypy (likeness to L1 speaker norms) of the associations they produce.

Measures of WA stereotypy are dependent on norms lists – lists of responses given to cue words by particular participant groups, ranked by response frequency. These can be used to investigate properties of particular words (e.g. *black* is considered to have a dominant primary response, because it so often elicits *white*), or of specific populations (based on, for example, age, gender, occupation, education, cultural background). Norms lists commonly used in applied linguistics WA research include

- the Postman and Keppel collection (1970)
- the Edinburgh Associative Thesaurus (Kiss et al. 1973)
- the University of Florida Free Association Norms (D. L. Nelson, McEvoy, & Schreiber 2004)
- the Small World of Words Norms lists in English and (so far) eleven other languages (databases for most of the latter are rather small at the time of writing, but the collection project is still live; De Deyne et al. 2018).

The predominance of English lists here prompts us to note two things: i) valid, peer-reviewed WA research in languages other than English is relatively hard to find, and ii) if Meara is correct in his suggestion that English has 'particularly high levels of stereotypy compared to other languages' (1980: 234), then norms lists in English are likely to be more reliably indicative of a population's response patterns than those in other languages. A norms list, much like a language corpus, is only as useful as its provenance and its fit with the target

research question. Fitzpatrick et. al. (2015) demonstrate that researchers' choice of norms list can influence stereotypy-related WA scores. Their study focused on (L1) WA responses to 100 cues by participants in two distinct age cohorts: 16-year olds and over 65s. Participants within each cohort were split into two experimental groups (16A, 16B, 65+A, 65+B) in a manner that matched exactly for age (since the participants were twin pairs), and a 'norms' list was compiled for each of the four groups. Each participant's response set was then scored against each of the four norms lists, with a point awarded for every response that appeared at the top of the norms list. Thus 4 separate norms list scores were generated for each participant. The mean score from the 'same age group' norms was significantly higher (26 out of 100) than those calculated from the two 'other age group' norms (19 and 19.2 out of 100), indicating that norms list data is influenced by age (or generation). Even before that study, many researchers had chosen to create their own bespoke lists to match the characteristics of their target population, but the Fitzpatrick et al. findings offer explicit evidence that norms list selection can critically affect stereotypy scores. This contributes to uncertainty about the validity of the 'L1 speaker norm' construct, as discussed below (see also Zhang & Koda 2017).

Insofar as the performance of an L1 speaker can be considered a benchmark against which to measure L2 proficiency, the expectation underpinning this stereotypy research - that the more proficient the learner, the closer their WA responses will be to the norms of L1 speakers of their target language – seems, on the face of it, a reasonable one. Two developmental considerations seem to support this: first, as proficiency progresses, responses are likely to reflect knowledge of multiple meanings and connotations of a word; second, in cases where response norms are translation equivalents across languages, the production of an L1 speaker-like response indicates that the learner has acquired the appropriate item in their L2. Nevertheless, despite promising early forays into this line of enquiry (Meara 1978;

Randall 1980), researchers have been confounded in attempts to find predictive relationships between L1 speaker-like responses and L2 proficiency. Wondering whether this problem is related to a WA protocol issue, many have tried adjusting various features of the task design, including:

- WA cue words – attention has been given to word class, frequency, propensity to elicit strong dominant responses (or not), and the number of cues used;
- Instruction – the number of responses requested for each cue has been varied;
- Scoring system – options include awarding 'stereotypy' points (a) for any response in the top e.g. 3 in the norms list; (b) for the percentage of the norming population giving the response; (c) according to the ranking of the response on the norms list; (d) for any response that appears anywhere in the norms list; (e) for a response that is the dominant response on the norms list (list adapted from Fitzpatrick et al. 2015: 33). In much of the literature, (b) and (c) are referred to as WEIGHTED STEREOTYPY SCORES, and the others as UNWEIGHTED.

In a proof of concept paper intended to establish a research tool for measuring learner proficiency, Schmitt (1998a) uses L1 speaker data to address problems he sees as inherent in WA protocols to date. He considers that i) the conventional requirement of one WA response per cue is inadequate for demonstrating word knowledge, and ii) production of a common (i.e. frequently produced) L1 speaker response might indicate a different (higher?) degree of word knowledge than production of an uncommon/idiosyncratic L1 speaker response.

Schmitt's proposed solution was to ask for three responses per cue, and from these to compile a norms list for 17 words from 100 L1 English speakers. He was concerned about variation in response dominance of cue words (some cues attracted the same response more often than others), and devised a complex graded score system to account for this, whereby a 'maximum' notional score was calculated for each cue. Participants' scores were calculated

by adding up the number of norms contributors who had given each of the participant's responses, and expressing this as a proportion of the maximum possible score. Schmitt calculates a mean proportion score for L1 speakers, but finds it challenging to identify a 'threshold' score at which learner responses become native-like; he resolves this by identifying four calculable levels of 'nativelikeness' in response behaviour. In a follow-up study (Schmitt 1998b) he tracks development of learner vocabulary using this scoring method, but results suggest backsliding and lack of progression (towards L1 norms), indicating that the method is still not fit for purpose. However, Schmitt (1998a) provided a useful platform for subsequent WA research, articulating clearly a number of WA characteristics that have subsequently been taken up for examination: variability of cue words' propensity to attract dominant responses, variability in L1 speaker responses, and the elusiveness of a conclusive definition of 'native-like response'.

Taking on board the problem of variable cue word behaviour, Wolter (2002) selected cues whose responses were heavily weighted towards the first three responses in the norms list data (i.e. a high proportion of respondents gave at least one of those three responses). He too asked for three responses to each cue. For his learner and L1 speaker participants he calculated weighted scores (accounting for the number of norms list contributors who gave that response) and unweighted scores (a point for any response also on the norms list). Although group means saw L1 speakers score higher than learners, some L1 participants scored below the learner mean. Correlations with a proficiency measure were moderate (less than 0.5 for both scoring methods). Like Schmitt, Wolter considers WA still to hold promise as a proficiency measure, and calls for three revisions to WA protocols in order for their potential to be fulfilled:

- avoid cue words that are delexicalised (i.e. have little inherent meaning, and are usually used with a noun phrase; e.g. make, get) or that elicit personal experience (e.g. travel);
- exclude from analysis the lowest scoring response to each cue word, and/or prompt words that tend to produce low scores; this would reduce the confounding effects of individualized responses;
- use a more fitting proficiency comparison measure.

Together with Zareva, Wolter returns to the challenge of WA analyses (Zareva & Wolter 2012) with an evaluation of three measures: associative commonality, lexico-syntactic patterns, and collocative analysis. The first of these is relevant to this section of our paper, addressing the question 'At the higher levels of proficiency…to what extent do learners' primary responses to familiar vocabulary correspond with native speakers' associations?' (2012: 47). Using a principled methodology, the authors categorised responses of L1 speakers, advanced and intermediate learners against L1 speaker norms, as 'nativelike common, nativelike idiosyncratic, and non-native like idiosyncratic' (2012: 51). They found significant differences between scores of L1 speakers and learner groups, but no difference between the intermediate and advanced learners; both produced a higher proportion of idiosyncratic than common, nativelike responses. The authors conclude that this method of scoring was not particularly promising. An earlier study by Zareva and colleagues (Zareva 2005; Zareva et al. 2005) had also found that comparison with L1-derived norms could distinguish between response data from learner and L1 participants. However, other ways of measuring the response data (number of associations generated, within-group consistency of associations) proved more sensitive to differences in proficiency than did L1 stereotypy. They suggest that extralinguistic factors, including cultural drivers, affect associative links, and that the nativelikeness score cannot therefore be seen as a clean reflection of proficiency.

While the above studies focused on scoring protocols as the key to extracting proficiency information from WA responses, other studies focused on cue word selection and norms list compilation. Munby's work is a key contribution here (2011; Fitzpatrick & Munby 2014; Munby 2018). His starting point was a scrutiny of Kruse et al. (1987), who had investigated the viability of WA as a measure of proficiency, comparing responses of 15 Dutch learners of English with those of 7 L1 speakers, using a multiple response protocol: participants provided up to 12 responses to each of 10 cue words. Kruse et al. used three measures: number of responses given; weighted stereotypy; and non-weighted stereotypy (using Postman & Keppel 1970). No strong correlation was found between any of these scores and proficiency measures (the highest was r=0.576), nor did they find any clear difference between scores of learners and the L1 speaker control group. The authors conclude that 'contrary to the expectations raised by earlier studies, we find that word association tests do not show much promise for the specific role created for them in L2 research' (1987: 153). This relatively small study seemed to temporarily stop proficiency-related WA research in its tracks (Meara 2009: xi); little research was conducted on this topic for a decade following its publication.

Scrutinising the methodological detail in Kruse et al. (1987), Fitzpatrick & Munby (2014) note as potentially problematic i) the selection of cue words, and ii) the scoring of multiple responses against norms lists that had been created from single responses. They set out to create a new version of the Kruse et al. test, using 10 cue words painstakingly selected for maximum sensitivity to proficiency, and creating a bespoke multiple-response norms list from 114 L1 speakers of English. 71 Japanese learners of English gave up to 12 responses to the new and the original Kruse et al. cues, and for each set of cues, an unweighted stereotypy score (using the Postman Keppel norms lists for the original Kruse et al. cues, and the bespoke norms list for the new cues) was compared with performance on three proficiency

tests. Correlations between proficiency tests and scores from the original cues were similar to those reported in Kruse et al., but response stereotypy for the authors' new cue set yielded a correlation of .7 ($p<.01$) with TOEIC scores.

Munby's next step was to question the use of L1 speaker norms lists in these studies. He and colleagues (Racine, Higginbotham & Munby 2014) make a strong argument against the use of L1 speaker norms as a benchmark for learners' WA performance, on the basis that i) L1 speaker responses are not homogeneous; ii) socio-cultural and demographic differences between L1 speakers and learners hinder linguistic comparability (this echoes Zareva et al.'s (2005) concerns about extralinguistic factors); and iii) such measures rest on the assumption that an L1-like variety of English is the learner's goal. Munby (2018) addresses this by compiling two separate norms lists for 50 cues: one from L1 English speakers (n=114) and one from L1 Japanese speakers who were highly proficient users of English (n=114). 82 L1 Japanese learners of English then completed the WA task, giving multiple responses to each of 50 cues. Their responses were awarded two scores, according to whether they were on i) the L1 and ii) the L2 norms lists. Learners scored significantly higher on the L2 than the L1 norms lists; Munby suggests this is perhaps because of shared cultural background. However, the marginally stronger – though still moderate – correlations with proficiency came from the L1 norms scores. Munby suggests this could be due to a native variety bent in proficiency tests, or to the fact that both proficiency and L1-like response behaviour are positively influenced by increased exposure to L1 English input.

It is our assessment that despite the sophisticated and dogged attempts by these researchers to find a method of stereotypy scoring that can reliably evaluate learner proficiency or word knowledge, this is likely to be an unattainable goal (though further calibrations of cue words and scoring protocols may chip away at statistical findings). We suggest two fundamental reasons for this. The first is a paradox connected with language

development: the more proficient a learner becomes, the more lexical items are available as potential responses in a WA task. However, the more proficient s/he becomes, the more honed and precise word selection can be, meaning fewer items present themselves as candidates for a WA response. It is possible that these things affect learners at different stages of development (vocabulary grows, then becomes more honed); production of native-like responses is therefore unlikely to operate on a steady cline.

The second reason for our assertion that stereotypy-based evaluations of L2 proficiency are unlikely to succeed is derived from usage-based (UB) theories of language development (e.g. Bybee & Beckner 2010; Bøyum 2016; Ellis, Römer, & O'Donnell 2016; Thwaites 2019, and see Section 5.3). These theories assert that an individual's linguistic system is the result of their unique experience with language. If this is the case, then it follows that the WA responses of learners, whose L2 experiences are often largely based on tightly controlled snippets of language found in classroom contexts, are unlikely to reflect the rich and diverse linguistic experience of an L1 speaker of the same language.

The next area of WA research we consider avoids managing unwieldy and unpredictable learner data sets, by examining not the associations a participant *produces*, but the associations s/he *recognises* from a restricted set of options.

3.3 Word associates format (WAF): testing depth of word knowledge

Whereas the stereotypy research strand examines WA responses generated by participants, the WAF presents participants with a set of predetermined items, and requires them to identify which are associated with the stimulus word. In this sense the WAF is rather different from WA models we focus on elsewhere in this paper: it does not ask for words to be produced, but rather for connections between words to be identified. This means that item

cues, associates and distractors can be manipulated according to research or testing aims. In terms of word knowledge, the focus is firmly on the stimulus word – it is the depth of knowledge of that word that is being investigated. Target associates are typically selected on the basis that they have collocational or definitional relationships with the stimulus, and so this kind of task can explicitly target aspects of word knowledge relating to meaning and syntactic relationships, and to polysemous uses of a target item. Because the WAF has verifiably 'correct' answers, and because teachers/testers can select target items according to curriculum needs, the WAF is useful to practitioners as well as to researchers, and is often proposed as a DEPTH OF (WORD) KNOWLEDGE test. Zhang & Koda (2017) offer a comprehensive review of WAF, and in particular its validity as a research tool.

The WAF was pioneered by Read (1993; 1995; 1998; 2000: 178–187), and he has worked systematically to develop it into an effective tool for assessing learners' knowledge of the meaning(s) and collocations of vocabulary items. In the last twenty years the WAF has been taken up by other researchers, who have further refined it, and applied it to data from a variety of languages and age groups.

In the original version of Read's test (1993), each stimulus word was presented alongside eight associates and distractors. The test-taker's task was to identify the associates, which could be 'paradigmatic' (defined by Read as wholly or partially synonymous), 'syntagmatic' (collocates), or ANALYTIC, which Read defines as representing one aspect or component of the target word, and likely to form part of its dictionary definition (1993: 181). For the example stimulus word *edit*, the associates/distractors are *arithmetic, film, pole, publishing, revise, risk, surface, text*. The intended correct associates are *revise* (paradigmatic), *film, text* (syntagmatic) and *publishing* (analytic).

Trials found a strong correlation with a word definition test, and Rasch analysis of test items indicated a good general level of reliability. However, post-task interviews revealed

several issues which led Read to make a number of revisions to the format. One change of particular significance was the presentation of the eight associates and distractors in two distinct sets – paradigmatic and syntagmatic. Participants selected the 4 target associates (these were not necessarily distributed evenly across the two categories). Read's walk-through account of this process meticulously explains the complex set of decisions involved in test development (2000: 180–187).

While refinements continue to be made to the WAF, it is important to note that studies have generally shown that scores on WAF tests can predict performance on linguistic tasks such as reading comprehension; the Depth of Vocabulary Knowledge (DVK) studies are important contributions in this regard (Qian 2002; Qian & Schedl 2004). In addition, Schmitt et al. (2011) found that WAF-style tests correlate closely with measures of depth of word knowledge derived from post-task interviews. These correlations, and those between the WAF and proficiency measures reported by Read, Qian, and Qian & Schedl are stronger, on the whole, than those obtained in the stereotypy studies reported in the previous section, and might indicate a more straightforward relationship between the two, or at least one with fewer confounding variables at play. We note, however, Horiba's (2012) finding that for Korean (but not for Chinese) L1 learners of Japanese, scores on the WAF test did not explain any additional variance on reading comprehension tasks than that which was explained by vocabulary breadth tests. For further discussion of this and the possible interaction with orthography it implies, see Zhang & Koda (2017).

Around the same time as Qian developed the DVK test, Greidanus & Nienhuis (2001) created a WAF test for Dutch L1 learners of French, and reinstated the analytic set of associates/distractors originally proposed by Read. Greidanus & Nienhuis focus on the nature of the distractors, comparing performance on items with semantically related and semantically unrelated distractors. An example of the former is the item Fr. *rive* 'bank', with

associates/distractors *artificial* 'artificial', *bord* 'edge', *côté* 'side', *fleuve* 'river', *gauche* 'left', *vague* 'wave'. Intended correct associates are *bord* (paradigmatic), *gauche* (syntagmatic) and *fleuve* (analytic). Their participants found the items with semantically related distractors challenging, and the authors conclude that these are therefore the appropriate kind to use; they also make guessing harder. Schmitt et al. (2011) make a similar comparison - between use of meaning-related and non-related distractors – and find an interaction with the number of options participants had to select from. Zhang & Koda conclude that the impact of distractor type should be considered alongside other influencing variables (2017: 14). These findings and their interpretation give us pause, because in any language encounter, including a WAF task, we are driven to make meaning, and in order to suppress the sense of a connection between, for example, *rive* 'bank' and *vague* 'wave' (we might visualise a wave lapping against a riverbank), it is perhaps necessary to employ analytic, metalinguistic skills. This is not in itself problematic, but it makes it difficult to be sure about whether knowledge *of* (a) language or knowledge *about* language is being tested.

Dronjic & Helms-Park also express concern about the type of knowledge being targeted in WAF tests; they consider the DVK to be a 'highly metalinguistic test' (2014: 211). Their paper resonates with the introduction to this State of the Art piece, by calling for lexical acquisition studies to be better informed by psycholinguistic research. Specifically, they argue that in order to understand how WA tasks work, more information is needed about i) lexical links in stable (L1) and developing (L2) networks; and ii) the nature of lexical activation in a WA context. On the basis that any test used with L2 learners should be able to produce consistent responses from L1 users, they analyse WAF data produced by L1 English undergraduates. Even in this relatively homogeneous group, they find high levels of variation in the identification of associations, particularly syntagmatic ones. They conclude that this lack of consistency in L1 data, together with concerns about differences between lexical

activation in WA tasks and in natural speech production, means that WA tasks are unfit for the evaluation of learner word knowledge, and should be used for exploratory research purposes only.

Mulder & Hulstijn (2011) conduct such an exploratory investigation in their critical examination of the notion of *native speaker proficiency*. They use an abridged WAF test along with a series of other tasks, to examine lexical knowledge and availability in adult L1 speakers of Dutch in three age ranges (18-35, 36-50, 51-76). Participants were presented with just one associate and one distractor for each target item, and were required to choose one. The researchers recorded an additional measure, response time, and found that older participants responded more slowly, but with equivalent accuracy to the younger participants. They conclude with cautionary remarks about the use of L1 speaker benchmarks in L2 studies, noting the variability in L1 speaker performance on the tests they administered.

Response time data is also used in Cremer & Schoonen's (2013) study of semantic knowledge in bilingual and monolingual children. They measure both response time and accuracy in a WAF task, and compare scores with reading comprehension performance. They identify a difference between LEXICAL AVAILABILITY (word knowledge) and LEXICAL ACCESSIBILITY (fluency) and find that accessibility accounts for some variance in reading comprehension that is not accounted for by word knowledge. Zhang & Koda conclude from these findings that omission of accessibility measures in WAF tests might underplay the predictive power of WAF tests (2017: 25).

As exemplified in Cremer & Schoonen (2013), the operational and ecological advantages of WAF-based tests – taking the test is not onerous and does not require language production but merely the circling of words – makes it particularly suitable for use with young participants. Schoonen & Verhallen (2008) capitalised on this, using a WAF test for 9 and 11 year old Dutch L1 or L2 speakers, in order to assess their vocabulary knowledge.

They asked participants to select from six items the three associates that ‘fit with the stimulus best’ or ‘always go together’ with the stimulus (2008: 218). They give the example of their stimulus item *banana* and associate/distractors *fruit, peel, yellow, slip, monkey, nice* (translated from Dutch); the first three are the most fitting associates. The authors find that performance on this test correlates strongly with that on a definition task. The reliability and validity analyses conducted in this study are sophisticated and the authors conclude that their results can guide further investigation into depth of word knowledge in children.

The final study to mention in this section, and one we will return to later, is Henriksen’s (2008) data-rich study of the production and identification of associations in two languages (Danish L1 and English L2) by participants representing three age groups: 13-14, 16-17 and early 20s. Her version of the WAF (she calls it the “word connection” task) is tied more closely to conventional WA work than those described above, in that her associates/distractors are all drawn from existing WA norms lists. For each of her 24 stimulus items (high frequency adjectives from the Kent Rosanoff lists), ten potential associates are provided: five are the most common responses to the stimulus item according to norms list data, and five are only given by one respondent in the norms lists – but are still semantically related to the stimulus. So, for the example stimulus word *cold*, the following possible associates are given: *war, water, frost, hand, hot, warm, snow, pain, winter, ice*. The instruction is to indicate the ‘five words they consider to be most strongly related to the stimulus word’. The author suggests that this gives participants ‘greater opportunity to draw on both their conceptual …. and their meta-semantic knowledge’ than the equivalent productive (conventional) WA task (2008: 43); that productive task, using the same cues as the word connection task, is discussed in Section 4. Henriksen’s multi-layered study offers comparison potential across several dimensions: participant-wise, there is an L1 (English) control group, and three groups of Danish L1 students at different ages (therefore educational

levels) and stages of L2 (English) development; language-wise, equivalent data is collected from the students in Danish and English; task-wise, participants not only undertake a conventional WA task and the "word connection" task, but also essay writing and a lexical inferencing task (reported by Albrechtsen & Haastrup respectively, in the same 2008 volume). This makes for rich pickings for WA researchers, and while the research reported in the 2008 volume likely does not exhaust the data's potential, some insightful observations and careful justifications emerge, not least the following from Henriksen, which relates to questions about L1 speaker benchmarks that arise elsewhere in the current paper: 'One could ask why the ability to identify or supply a native-like response is seen as a hallmark of a well-developed mental lexicon. The underlying assumption is that native-like associational patterns will reflect the types of conventional access routes or lexical pathways in the mental network available to a fully competent language user in communicative situations' (2008: 44).

Juxtaposing the various operationalisations of the WAF discussed above enables us to evaluate its contribution to the WA research field. Superficially, the precise instruction and restricted response options lead us to suppose that it is more straightforward to identify the construct being tested here than in the more unwieldy conventional WA tasks, where the potential response pool is enormous. For each target word, manipulation of associates can reveal test takers' knowledge of specific collocations or multiple meanings of polysemous words, and so on. However, some studies suggest that different knowledge constructs are targeted in the identification of syntagmatic and paradigmatic (and perhaps analytic) associations (e.g. Greidanus & Nienhuis 2001; Batty 2012; Dronjic & Helms-Park 2014; see also Zhang & Koda 2017). This links with considerations of RESPONSE TYPES addressed in the following section, and rather than being seen as intractably problematic, we suggest that the designed distinction between association types in WAF tests might offer a potential route to

unpicking some of the conflicts and white noise associated with categorisation of responses in the research discussed in the following section.

We also detect in this WAF research a tension between constructs that goes beyond types of word knowledge. This tension is between i) intuitive judgements, especially L1 speaker intuition, which has been shown to be unreliable, for example by McGee (2009); ii) knowledge of individual vocabulary items (the stimuli); and iii) metalinguistic knowledge, which Dronjic & Helms-Park claim is key to successful completion of the WAF. Again, the interplay of these different dimensions in WAF task completion might be exploited, perhaps in addressing a question that has emerged quite strongly from this section: that of the extent to which L1 speaker data provides a useful or even valid benchmark for learner WA tasks.

Finally, we note that this research strand is less dominated by English than those described in the previous section; it is likely that the 'high levels of stereotypy' offered by English over other languages (Meara 1980) are not so relevant to the format of the restricted choice Word Associates Format, which can be manipulated to elicit different layers of knowledge.

4. Using word associations to investigate word storage and retrieval – category approaches

The previous section focused on the typicality of responses to a cue, and what responses can tell us about knowledge of the cue word. The studies in this section are concerned with the linguistic relationship between cue and response, and use this information to investigate the way words are stored and retrieved for use. The principle driving this strand of WA research is that responses can be assigned to discrete categories of association, and that the resulting 'scores' can be related to other variables, such as the respondent's language proficiency,

background, whether they are using their first or second language, or to characteristics of the cue words.

Although WA responses had been analysed by category in some earlier research, it was Osgood et al. who in 1954 formally articulated a system that came to dominate studies of L1 and L2 language development in the following decades (and which we have already referenced in this paper). Criticising previous WA classifications as 'an unsystematic mixture of semantic, psychological and linguistic criteria' (1954: 115), they suggested a linguistic system based on two classes: paradigmatic – if cue and response can be substituted within the same syntactic frame – and syntagmatic, where one word is followed immediately by the other in usage (even where both are the same word class), with the likely addition of a 'phonetic similarity' category, later referred to as CLANG (1954: 116). Their predictions that this or any system is unlikely to account for *all* responses, that refinements are likely to be needed, and that objective assessment will be problematic, all foreshadow challenges that plague later studies, but the paradigmatic/syntagmatic approach to analysis persists in WA research. We have already noted its use in the WAF investigations of word knowledge (see Section 3.3); in this section we take it as a starting point for examining lexical organisation and processing.

A series of influential studies in the 1960s applied the paradigmatic/syntagmatic system to WA data from children (e.g. Brown & Berko 1960; Ervin 1961; Entwisle et al. 1964). A difference was detected between WA behaviour of older and younger children, with older groups producing a higher proportion of paradigmatic responses and younger groups a comparatively high proportion of syntagmatic and clang responses. A similar pattern had emerged in comparisons of child and adult data from much earlier and less cited studies, as reported in Woodworth (1938). Table 1, from Woodworth, illustrates these: *eat* and *hole* would be classed as syntagmatic responses; *chair* and *shallow* as paradigmatic.

Table 1 Comparison of child and adult WA data from Woodrow and Lowell (1916) and O'Connor (1928) respectively, cited in Woodworth (1938: 346)

Cue	Response	1000 children	1000 adults
table	eat	358	63
	chair	24	274
deep	hole	257	32
	shallow	6	180

Other than a brief note of similarities between L1 and L2 acquisition in Brown & Berko (1960: 4), there is little mention of second language acquisition in this child language development literature. However, the apparent identification of a shift in L1 development from clang to syntagmatic to paradigmatic associational links inspired researchers to look for an equivalent developmental shift in L2 learners, as proficiency increased. Of the many subsequent L2 studies, few claim to have detected a clear syntagmatic-paradigmatic shift. Scholars have differed in the way they have defined the paradigmatic, syntagmatic and clang categories, and many have added extra, or alternative, categories. This, alongside other methodological differences (see Section 4.2), makes comparability of findings difficult. Nevertheless, some persistent claims about WA behaviour have emerged, and are summarised in the following section.

4.1 Main claims in category-based word association research

There has been a steady momentum to this strand of research in the last two decades, and despite persistent confounds, some broad patterns of WA behaviour can be detected, and are summarised below. The weight of evidence behind these varies, and occasionally findings are contradictory (marked ! below). It will be noted that in addition to the *paradigmatic*, *syntagmatic* and *clang* categories mentioned above, the terms MEANING-BASED, POSITION-

BASED and FORM-BASED are used as category labels. These are sometimes considered more user-friendly categories, and are often subdivided further. Specific categorisation systems are discussed further in Section 4.2.

Indicative findings from category-based WA research, with examples of studies reporting them, include:

- In the majority of studies, paradigmatic or meaning-based responses are most prevalent, and clang or form-based responses least prevalent.
- The likelihood that a participant will produce a paradigmatic response or a meaning-based response is increased:
    - if they know the cue word well enough to use it in a sentence (Wolter 2001)
    - if the (L2) cue word is relatively familiar (Söderman 1993; Wolter 2001)
    - if they are an expert user (L1 or advanced L2) of the language of the WA task (Zareva 2007; Jiang & Zhang 2019)
    - the more proficient in L2 they are (Söderman 1993; Orita 2002; Zareva 2007; Zareva & Wolter 2012; Khazaeenezhad & Alibabaee 2013)
    - if the cue words are nouns (Nissen & Henriksen 2006)
    - the older they are (Namei 2004)
    - if they are adult (Cremer et al. 2011)
    - if they are a heritage L2 speaker (Kim 2013)
- The likelihood that a participant will produce a clang or orthographic/phonological response is increased:
    - the younger they are (Namei 2004)
    - if they are using their L2 (Wolter 2001; Namei 2004; Fitzpatrick 2006; Norrby & Håkansson 2007; Hui 2011; Jiang & Zhang 2019)

  - if they are learning L2 outside the target language environment (Håkansson & Norrby 2010)
  - the less proficient in L2 they are (Söderman 1993; Orita 2002; Khazaeenezhad & Alibabaee 2013)
  - if the cue word is relatively unfamiliar or newly-acquired (Söderman 1993; Wolter 2001)
- The likelihood that a participant will produce a syntagmatic/collocational/position-based response is increased:
  - ! if they are using their L2 (Norrby & Håkansson 2007; Zareva 2007; Håkansson & Norrby 2010)
  - ! if they are using their L1 (Fitzpatrick 2006; Fitzpatrick & Izura 2011)
  - if they are learning their L2 as a foreign rather than second language (Norrby & Håkansson 2007)
  - the less proficient they are (Söderman 1993; Orita 2002; Zareva 2007; Zareva & Wolter 2012; Khazaeenezhad & Alibabaee 2013)
  - if the cue words are adjectives (Nissen & Henriksen 2006)
- The likelihood that a participant will produce a synonym response is increased:
  - if they are using their L1 (Fitzpatrick 2006; Fitzpatrick & Izura 2011)
  - the more proficient they are (Khazaeenezhad & Alibabaee 2013)
- The likelihood that a participant will produce a loose conceptual association response is increased:
  - ! if they are using their L2 (Fitzpatrick 2006; but not supported by Kim, 2013)
- The likelihood that a participant will produce a quality-based association response is increased:
  - if they are using their L1 (Kim 2013)

These findings must be considered in context, of course; results that feed into the broad statements above are nuanced according to specific proficiency levels, languages investigated, cue words used and so on, and scrutiny of these and other variables is necessary to avoid further research being driven down cul-de-sacs by false assumptions. Additionally, some of the studies conflate variables that others examine separately; age and proficiency in Orita (2002) and Namei (2004) for example. Zareva's (2007) identification of items unknown to participants in order that the confounding effect of these might be excluded from analysis is a good but rare example of an attempt to manage data so as to eliminate confounds. Henriksen (2008) implicitly conveys a warning against assumptions of linear or binary models of WA behaviour, in the finding that learners simultaneously produce more 'canonical' (prototypical) and more low-frequency responses as development progresses. Crucially, it should be noted that there are few, if any, precise replications of the studies considered here, and in our view these are crucial in order to weed out true findings from context-dependent ones.

In the next section we unpack some of the methodological differences, challenges and approaches found in this research strand.

4.2 Word association response categorisation: methodological approaches

Scrutiny of the studies in this research strand reveal differences on multiple dimensions which, as noted above, make it difficult to compare findings across studies. Differences include:

- the precise definition of the paradigmatic/syntagmatic/clang categories
- the number of cue words used and the number of responses elicited for each cue

- whether/how the data is prepared for analysis (e.g. Zareva & Wolter 2012 operate detailed lemmatisation criteria)
- the theoretical models (or absence of) motivating the study
- whether comparisons are between-subject or within-subject, and if the latter whether they are longitudinal, comparing participant data as proficiency progresses (e.g. Fitzpatrick 2012); or cross-sectional, comparing participants' L1 and L2 behaviour (e.g. Politzer 1978; Nissen & Henriksen 2006)
- whether between-subject comparisons are made on the basis of L2 proficiency level (e.g. Orita 2002), identity as learner or L1 speaker (e.g. Wolter 2001), or some other criterion
- whether the focus is on the holistic organisation of the lexicon, or on individual lexical items (e.g. Söderman 1993; Wolter 2001)
- which languages are investigated, and the extent to which features of a specific language affect response behaviour (see Kim 2013, on Korean)

To further confound matters, the reach of this research strand has been (and continues to be) reshaped in terms of both methodology and the target variables. The range of target variables has in some cases been extended (for example to include learning context: foreign language, second language, or heritage language), and in others has been reduced to a focus on particular developmental or organisational features (such as the influence of word form). Most notably, researchers have developed more fine grained, or transparent, or fit-for-(specific)-purpose categorisation systems, reflecting the theoretical drivers of their research. In Table 2 we present an overview of some of the most salient and/or cited studies from the last 20 years, organised according to the approach taken to classification of responses.

Table 2 WA studies using category approaches: a methodological overview (post-2000)

Paper	Cues	Responses	Classification Approach	Categories used
Wolter (2001)	90	1	broadly paradigmatic, syntagmatic, clang	4: paradigmatic, syntagmatic, clang, no response
Orita (2002)	60	1		4: paradigmatic, syntagmatic, phonological, other
Namei (2004)	100x2	1		3: paradigmatic, syntagmatic, clang
Nissen & Henriksen (2006)	90x2	2		4: paradigmatic, syntagmatic, phonological, other
Zareva (2007)	73	≤ 3		3: paradigmatic, syntagmatic, phonological (commonality measures also used)
Yu, Xu & Sun (2011)	30	1		4: phonological, syntagmatic, paradigmatic, other
Zareva & Wolter (2012), analysis 1 (of 2)	36	1		2: paradigmatic, syntagmatic
Roux (2013), analysis 1 (of 2)	8	1		3: paradigmatic, syntagmatic, clang/phonological

Norrby & Håkansson (2007)	100x2	1	augmented paradigmatic, syntagmatic, clang	4: paradigmatic, syntagmatic, phonological, and translations
Håkansson & Norrby (2010)	100	1		4: paradigmatic, syntagmatic, phonological, and translations
Khazaeenezhad & Alibabaee (2013)	20	1		4 main, 11 sub or sub-sub categories: paradigmatic (co-ordination (complementary, gradable, converses, mutual incompatibles), hyponymy/ hypernymy synonymy); syntagmatic (lexical, grammatical, restricted collocations); phonological/orthographic; encyclopaedic
Fitzpatrick (2006)	60	1	version of Fitzpatrick (2006, 2007)	4 main, 17 sub-categories: meaning-based (defining synonym, specific synonym, hierarchical/lexical set, quality, context, conceptual); position-based (consecutive xy, consecutive yx, phrasal xy, phrasal yx, different word class collocation); form-based (derivational, inflectional, similar in form only, similar form association); erratic (false cognate, no link)
Fitzpatrick (2007)	100	1		4 main, 9 sub-categories: meaning-based (defining synonym, specific synonym, lexical set, conceptual); position-based (consecutive xy, consecutive yx, other collocation); form-based (change of affix, similar in form only); erratic (no link)
Cremer et al. (2011)	118	1		4 main, 17 sub-categories: Meaning-related (coordinate, subordinate, superordinate, antonym, partonym (part–whole), partonym (whole–part), context-independent, goal/target, synonym); Indirect Meaning-related (subjective association, composite word, context-dependent; Form-based association (change of affix, similar form); Other (non-classifiable, repetition, no response)
Fitzpatrick & Izura (2011)	190	1		equivalent meaning, non-equivalent meaning, collocation, form, form+meaning, meaning+collocation

Hui (2011)	85	1		4 main, 16 sub-categories: Meaning-based (synonym, context related, personal feeling); position-based (consecutive xy, consecutive yx); form-based (derivation, changed part of speech, compounds, +/- one letter, similar in form not meaning, repetition), ; erratic (wrong derivation, illegal creation, spelling error, no relation, letter cluster)
Kim (2013)	10	3		4 main, 12 sub-categories: Meaning-based (definition/antonym/synonym, hierarchical/lexical set, quality, strong conceptual, loose conceptual); position-based (SV, OV, other relationship, collocation); form-based (derivational/inflectional affix, similar in form); erratic.
Roux (2013), analysis 2 (of 2)	8	1		4: meaning-based, position-based, form-based, erratic
Yokokawa et al. (2002)	20x2	n/a	other	6: antonym, syntagmatic, category-exemplar, exemplar-exemplar, synonym, other
Henriksen (2008)	48x2	2		10: canonical (top 5 responses in baseline data, if given by >10%) + high frequency item; canonical + low frequency item; non-canonical + high frequency; non-canonical + low frequency; chaining (related to previous responses rather than cue); form-related; translation; repetition (of cue); empty/blank; ragbag (indecipherable or impossible to categorise)
Li, Zhang & Wang (2010)	33x2	1		4 categories: thematic, taxonomic, other semantic, non-semantic
Zareva & Wolter (2012) analysis 2 (of 2)	36	1		2: collocational or not collocational (commonality measures also used)
Jiang & Zhang (2019)	74	1		4 main categories: morphological, semantic, form, other

Initial forays into the application of the paradigmatic/syntagmatic classification in L2 research were promising in that findings supported the hypothesis that the paradigmatic>syntagmatic shift can be detected in L2 as well as L1 development. Politzer's (1978) comparison of L1 (English) and L2 (French) WAs, though vulnerable to criticism for the ambitious interpretation of findings, is notable for its relative statistical power (n=203) and the alignment of findings with the L1 literature: he found proportionally more paradigmatic responses, and fewer clang responses, in the L1 than in the L2 task, and found that scores on a French grammar test correlated positively with the number of paradigmatic responses produced. Following Politzer, the paradigmatic/syntagmatic/clang categorisation was applied in a series of L2 studies, with the label 'phonological' sometimes used in place of 'clang'. However, as Osgood et al. (1954) had predicted, categorisation was not straightforward; some studies give detailed explanations of how the classification was operationalized (e.g. Namei 2004; Zareva 2007), some add an 'other' category (e.g. Orita 2002; Nissen & Henriksen 2006; Yu, Xu & Sun 2011), and several researchers explicitly set out the drawbacks of the paradigmatic/syntagmatic/clang classification. Weaknesses observed include:

- there is a lack of 'clear and objective criteria' for assigning responses to categories (Henriksen 2008: 46)
- the breadth of the categories may mask information within them (Fitzpatrick 2006: 126; Nissen & Henriksen 2006: 46)
- the paradigmatic/syntagmatic distinction 'is very difficult to work in practice, especially when you cannot refer back to the testee for elucidation' (Meara 1983: 30); categorisation often entails making assumptions about context or mental mapping (Roux 2013: 87)

- it can be difficult or impossible to assign an association to (only) one of paradigmatic or syntagmatic (see e.g. Maréchal 1995, cited in Singleton 1999: 234; Fitzpatrick 2006; Nissen & Henriksen 2006)
- hyponymous semantic relationships are typically paradigmatic, but are sometimes represented by words of different classes (Some categorisation systems explicitly address this – see Wolter 2001: 52 and Namei 2004: 372)
- depending on the way in which the category distinction is operationalized, some associations do not fit into any of them (for example idiosyncratic associations, which are often emotional or experiential; Henriksen 2008: 46)
- category of association depends in part on word class; 'nouns seem to be predominantly organised in paradigmatic relations, whereas verbs and adjectives are characterised by syntagmatic relations' (Nissen & Henriksen 2006: 46)

These observations have led to the augmentation of the paradigmatic/syntagmatic/clang system in some studies, and the application of alternative classification systems in others. Augmentations have included the addition of a translations category (Norrby & Håkansson 2007; Håkansson & Norrby 2010) and extensive use of sub-categories (e.g. Khazaeenezhad & Alibabaee 2013). Of the alternative categorisation systems to have been proposed, some are theoretically motivated and others driven by data characteristics.

The most widely used alternative classification systems somewhat echo the traditional paradigmatic/syntagmatic/clang approach, despite deriving from different theoretical frames. Fitzpatrick acknowledges that her broad Meaning/Position/Form distinction, for example, bears similarity to the traditional approach, despite being partly inspired by pedagogic word knowledge frameworks (2006: 130–131). The division of those main categories into sub-categories, though, deviates from previous approaches and goes some way towards

addressing problems noted above: category labels are reasonably transparent, and it is fine-grained enough to observe group differences that would be masked by the conventional three categories - for example within the Meaning category, L1 speakers are found to produce significantly more synonymous responses (e.g. *church>cathedral*), and fewer responses with loose conceptual links (e.g. *culture>cathedral*) than learners (2006: 132–133). Use of multiple categories results in the creation of a 'profile' for each dataset, consisting of the number/proportion of responses in each category. The statistical challenge of comparing profiles, rather than single scores, is addressed by applying proximity analyses (see Fitzpatrick 2007; 2009). The advantage of the profile approach is that it enables holistic comparisons that are inclusive of all data produced in a WA task. Findings have indicated that individuals' profiles are relatively stable across test times (Fitzpatrick 2007) and across languages, with a suggestion that an individual's L2 profile will move closer to their L1 profile as proficiency increases (Fitzpatrick 2009). There are difficulties inherent in Fitzpatrick's system too, though, including:

- having up to 17 sub-categories necessitates a relatively large data set from each participant, in order for responses to be distributed densely enough for robust statistical analysis
- as with the traditional approach, it can be difficult to make a call between categories for many responses, challenging rater reliability (e.g. Roux 2013: 87)
- related to this, the categories are such that some responses fall squarely into more than one (sub-)category – this is the case, for example, for some of Henriksen's (2008) 'canonical' responses, such as fork>knife (fitting both collocation and lexical set categories)
- as with the traditional approach, it is not always possible to identify the nature of the association without referring back to the respondent, and even time-consuming post

task interviews depend on participants' capacity to account for a spontaneous response (see Fitzpatrick 2006).

Fitzpatrick's system has been tweaked in subsequent attempts to address some of these weaknesses, and/or to match the system better to specific research questions. For example Hui (2011) adds the subcategory 'personal feelings', Kim (2013) distinguishes between subject-verb and object-verb collocations to better fit the language of enquiry (Korean), and Fitzpatrick (2007) minimises the dilution of data power across categories by reducing the number of subcategories from 17 to 9. Fitzpatrick and Izura (2011) include dual-link categories, such as 'meaning and collocation' (e.g. *rubbish>bin*) and 'form and meaning' (*hairdresser>hairdryer*). While the adaptations may have made the system more rater-friendly, the fact that the system has not been held steady across studies makes replication and comparison very difficult, and problems of rater confidence and theoretical heft remain. On the other hand, the apparent consistency of individual profiles are broadly consistent across test times and between languages is a positive indication of the system's validity (Fitzpatrick 2007; 2009).

Other recent studies have used rather different approaches, for example focussing on the particular kinds of association that are pertinent to their research question. Examples of this are Jiang & Zhang's focus on form (2019), Yokokawa et al.'s focus on semantics (2002), Cremer et al.'s interest in conceptual vs. semantic knowledge (2011), and Zareva & Wolter's concentration on collocation (2012); these are discussed further in the following section. Finally, several studies layer categorisation of responses with other measures: for example, Fitzpatrick & Izura (2011) include response time measures, and Henriksen (2008) incorporates FREQUENCY of the response item within her sophisticated measurement system.

4.3 Theoretical influences on categorisation methodology in word association

Meara (1983) lamented the lack of theoretical models applied to WA behaviour, and it is still the case that while most WA researchers are linguistically aware enough to associate the paradigmatic syntagmatic distinction with the work of Saussure, surprisingly few of them explicitly expose or explore the provenance and implications of these terms. Stronger adherence to the Saussurian classifications might in fact serve WA researchers well, given the clarity of theoretical distinction between the 'in praesentia' (evident) co-presence of 'combinations based on sequentiality' – i.e. syntagmatic relationships, and the 'in absentia' 'connexion in the brain' of words 'linked by meaning only' – what Saussure calls associative relationships, and which are later labelled paradigmatic (Saussure, in Bally & Sechehaye 1966). Saussure's considerations of these notions include a number of features that have not systematically been transferred to WA research despite being, in our view, highly relevant. For example, within the syntagmatic class he includes combinations of parts of words and compound words, as well as entire words, and associations can be either double ('based on form and meaning') or single, based 'on form or meaning alone' (1966: 124). Operationalisations of the paradigmatic/syntagmatic distinction in WA literature vary in their closeness to the Saussure conceptualisation, but most WA category systems, as we have seen, distinguish in some way between three relationship axes that are identifiable in Saussure's work -

- relationships based on meaning, similarity of signifier, substitution, mnemonic groupings, paradigm, mental association
- relationships based on combination, expressions, position (within utterance), interdependence of units within an utterance
- relationships based on word form and/or formation, phonology and/or orthography

The extent to which theoretical models of any kind are referred to in connection with categorisation systems varies considerably. Many (e.g. Roux 2013) take previous empirical studies as a starting point, with the implicit assumption that the systems used in these have theoretical validity. Others (e.g. Namei 2004) examine theoretical underpinnings in previous WA research, but do not connect these with the specific categorisation systems they or others use. For Wolter (2001) theoretical modelling drives the research questions; he challenges the prevalent belief that a syntagmatic>paradigmatic shift indicates development of the lexicon, and proposes an alternative model that considers the trajectories of individual words, and ascribes changes in WA behaviour to a shift from 'semantically meaningless…. to semantically meaningful responses' (2001: 63). What is known about a particular word also contributes to Henriksen's classifications (2008), and the various iterations of Fitzpatrick's (2006) system. Zareva & Wolter's (2012) consideration that there are multiple aspects of the learner's lexical organisation, and that WA methods might tap into these differently, might also be considered a learner- or data-driven account.

In contrast, theoretical models of lexical organisation or development explicitly drive the methodology of some WA studies. Jiang & Zhang (2019) explore the hypothesis that the L2 lexicon is more form-driven than the L1 lexicon, and consequently two of their four WA categories are 'morphological' and 'form-related'. Yokokawa et. al. (2002) use fine grained semantic categories to investigate semantic organisation of the lexicon. Zareva & Wolter's classification of +/- collocational, in the second analysis of their 2012 paper, explores claims that acquisition is collocation-driven (citing Wray 2002 and Hoey 2005).

Finally, the prevalence of English language focussed WA studies should not obscure the role of linguistic typology in classification systems; there is an element of language-specificity in the way participants respond to WA cues. The differences Fitzpatrick (2009) finds in the number of forward and reverse collocations produced in English and Welsh is

almost certainly caused by word order differences (Welsh is VSO with noun typically preceding adjective), and Kim (2013) re-designs her classification system specifically to accommodate the highly-inflected, SOV Korean. Typological differences are a particular challenge in studies comparing bilinguals' responses across two languages, and the degree to which the structures of the relevant languages are accounted for differs; in our view, too little attention is paid to this.

Examining the theoretical underpinnings and challenges of WA research can expose opportunities for future research as well as weaknesses in previous work. In the next section we identify and explore three themes that we believe are key to future development of WA research.

5. Key notions for future word association research

In the previous sections we have scrutinised research that uses WA to examine word knowledge (Section 3), and that uses WA to investigate the structure, storage and retrieval routes of the mental lexicon (Section 4). We also considered the scientific context from which WA research emerged, and the interplay of disciplines involved. In this final section we identify three recurrent themes in the WA literature that we consider to be crucial if the potential of this research field is to be fully realised, particularly in relation to L2 development. The themes can be summarised as network models; lexical variables; and psycholinguistic constructs, and we address them in turn below.

5.1 Mapping word association findings onto network models of the lexicon

In a recent overview chapter, De Deyne & Storms (2015) identify three approaches to researching WA: MICROSCOPIC, MESOSCOPIC, and MACROSCOPIC. The research discussed in

Sections 3 and 4 above largely takes a microscopic approach – it is concerned with relations between individual cues and responses. This approach can, as Wilks & Meara have pointed out, 'be criticised for attempting to generalize the features and properties of a large scale phenomenon – the mental lexicon – on the basis of small scale snapshots of that phenomenon' (2002: 306) .

Meso- and macroscopic approaches take a wider view. Researchers in these traditions assert that a key feature of WA research is its ability to yield structured patterns of connections between words. This strand of research was founded by Deese (1962; 1966), who studied patterns of association related to cue words from the Minnesota Norms list (Russell & Jenkins 1954). In his study, responses to those cues were then used as cues in subsequent rounds of WA. Responses to all of these words can then be tabulated to form a matrix revealing how frequently each word in the dataset yields each other word when presented as a cue. Using this approach, Deese revealed that WA responses cluster semantically. For example, *butterfly* yielded responses such as *moth, insect, blue* and *color*. After responses to these new words were collected, factor analysis revealed several clusters of mutual association. One, for example, related to winged creatures (*butterfly, moth, bird, wing*), while another centred around colours (*color, blue, yellow*).

Deese's findings led straightforwardly to the mesoscopic WA research tradition, which focuses on the hypothesis that the overlap in associative distribution of two words can be taken as a measure of the semantic similarity of those words. For example, since there is greater overlap in associative response distributions between *moth* and *butterfly* than between *colour* and *butterfly*, it can be asserted that the former pair are more similar than the latter. Results from studies investigating the capacity for WA data to generate semantic similarity ratings in this manner (e.g. Steyvers, Shiffrin, & Nelson 2004; Andrews, Vinson, & Vigliocco 2008; Van Rensbergen, De Deyne, & Storms 2016) have been impressive: such

ratings have been shown to correlate with human judgements of semantic similarity, and to explain variance on psycholinguistic tasks such as lexical decision (Steyvers et al. 2004; De Deyne, Navarro, & Storms 2012). These results suggest that networks built from WA data are not only suitable for use as simple estimates of semantic similarity for psycholinguistic experiments, but also that they capture important aspects of lexical processing.

Deese's work was important also in the development of network approaches to lexical research. This view proposes that word knowledge is organised in a massively interconnected, web-like structure (Wilks & Meara 2002; Aitchison 2012). Research in the macroscopic tradition has used WA data to build approximations of such networks, and to explore their implications for lexical development, attrition, and global structure. Although research into network structure oversimplifies actual lexical knowledge, it nevertheless offers a crucial link between the lexicon and lexical fluency, since, as Strogatz puts it, 'structure always affects function' (2001: 268).

An early example of macroscopic WA research is found in Kiss's (1968) application of graph theory to WA-derived networks. Graph theory is a way of describing networks, such that NODES (corresponding, in WA networks, to a single word) are connected by LINKS (also called ARCS or EDGES). In some models, these links are WEIGHTED. In WA-derived networks, weights are derived from the association strength between two words. Links can also be DIRECTED, meaning that each link travels in only one direction. This allows WA-derived networks to correspond to the directed nature of some WA pairs. For example, in a weighted, directed network, two links would join *writing* to *pen*, since both words yield the other in free association. However, since *writing* yields *pen* much more commonly than the reverse (according to the South Florida norms; D. L. Nelson et al. 2004), the link from *writing* to *pen* would be stronger than that of *pen* to *writing*.

A number of important measures can be derived from these networks, some focusing on individual nodes, and others measuring the properties of the network as a whole. An important measure of individual nodes is their DEGREE. This refers to the number of links a node possesses to other nodes in the network. In directed networks, which differentiate between incoming and outgoing links, the terms IN-DEGREE and OUT-DEGREE are used. In WA networks, in-degree refers to the number of cues to which word *A* is given as a response, while out-degree measures the number of different responses given to word *A* when it is presented as a cue. Continuing the example above, in the South Florida norms (D. L. Nelson et al. 2004), *writing* was given as a response to a total of 14 cues, yielding an in-degree of 14. When *writing* was presented as a cue, 21 response types[iii] were given as responses, resulting in an out-degree of 21[iv]. Words which possess many links to other words can be described as being central to the network. Some research suggests that associative network CENTRALITY may influence online lexical processing, since measures of associative centrality can explain unique variance in lexical processing tasks such as lexical decision, categorisation, and visual word recognition (Griffiths, Steyvers, & Firl 2007; Duñabeitia, Avilés, & Carreiras 2008).

Researchers have suggested that central words play a formative role in the growth of the first language lexicon. This hypothetically occurs through a process known as PREFERENTIAL ATTACHMENT (e.g. Steyvers & Tenenbaum 2005), whereby newly acquired nodes preferentially attach to nodes which are already central to the network. This leads to the prediction that frequent and early acquired words should become increasingly central, because they have the highest probability of attracting links to new nodes. Support for this theory has been provided by De Deyne & Storms (2008), who found that both word frequency and age of acquisition were correlated with centrality in a WA-derived network; and from simulations run by Meara (2007). One caveat to these findings is that, because few studies have utilized the most sophisticated of the available measures, it remains unclear

whether WA-derived centrality captures aspects of processing not explained by other measures (see e.g. De Deyne & Storms 2008; Yap et al. 2011). Nevertheless, the study of network centrality in L1 lexical development raises the possibility that this approach may shed light on some of the determinants of L2 lexical acquisition.

Lexical researchers have been particularly interested in SMALL WORLD networks (Milgram 1967; Watts & Strogatz 1998). This type of network displays several properties not found in random networks: sparseness (e.g. a network of 122,005 nodes, created from the WordNet database by Steyvers & Tenenbaum (2005: 51), had an average of only 1.6 links per node); short average path lengths (e.g. an average of just 10.56 links connected any two nodes in the same network); and high clustering, indicating the presence of highly interconnected neighbourhoods, as exemplified by the semantic clusters discovered by Deese (1966), and possessing relatively few links to other clusters. Networks such as these depend on 'hubs'[v] – a small number of words which have a very high degree, including connections between several neighbourhoods – to facilitate fast search and retrieval of items analogous to the fast semantic access of which humans are capable (De Deyne & Storms 2015: 477).

The studies reported above seem to confirm that WA-derived networks display small world properties. However, they all take a between-subject, cross-sectional approach, resulting in networks based on responses from a wide range of participants, whose individual response preferences are not known. As such, they may be best understood as reflecting broad possibilities regarding network structure, rather than as revealing the properties of any one person's actual associative network. A study by Morais, Olsson & Schooler (2013) took a different approach, exploring the structure of six associative networks each derived from WA responses given by one person. This study revealed marked variation in the size, connectedness, path length, and clustering of each network, suggesting clear individual variation in the properties of the systems or processes underlying WA. Importantly, however,

all of these networks demonstrated the same general small world structural properties, in spite of their surface-level differences. Related to this approach is work by Beckage, Smith, & Hills (2010) who, in a study of the L1 networks of young children, linked delayed emergence of these small world properties with later language processing difficulties. These studies suggest that small world network properties may be reflective of fully developed associative knowledge, and may also be related to online language processing. This is a potential area of interest for second language researchers: if the development of efficient networks of lexical knowledge can be demonstrated to be a precondition of verbal fluency, this would appear to prioritise research on how the emergence of such networks in L2 learners can be facilitated in the classroom.

The application of these approaches to L2 research has been hampered by methodological challenges. A central problem, which applies to several of the studies discussed below, is that L2 responses tend to be less consistent than those produced by L1 respondents (Meara 2009). This inconsistency has led some researchers to abandon the use of productive WA tasks for the construction of L2 associative networks, and instead develop receptive tasks in which participants identify associated words from lists of provided options (similar to the WAF tasks discussed in Section 3.3, but without the notion of correct/incorrect answers). Using this approach, Wilks and Meara, in a series of studies (Wilks & Meara 2002; 2007; Wilks, Meara, & Wolter 2005; Wilks 2009), suggest that L1 networks are significantly denser than those of L2 respondents, and they observe that increases in L2 network density are linked to L2 proficiency.

Meara (1996b; 2007) progresses the application of network theory to L2 lexical research in his discussion of the concept of depth of word knowledge. Rather than viewing this concept as a property of individual words (see Section 3), he argues for a network-based measurement of global depth of vocabulary knowledge which conceptualizes depth as the

total number of links within a learner's lexicon. However, he acknowledges that a sufficiently large number of test items and an appropriate test methodology are required in order to generate a meaningful indication of depth of vocabulary knowledge, and methodological problems have beset trials using receptive WA tests (e.g. Meara & Wolter 2004; Wilks & Meara 2007).

Another attempt to measure L2 network properties through receptive WA tasks was made by Schur (2007), who investigated the small-world properties of L2 networks through a task in which monolingual and bilingual (Hebrew/English and Chinese/English) participants identified associates from a list of 50 frequent English verbs. Although this methodology proved too limited to fully achieve its aims, it did reveal interesting differences in participant groups: the Chinese participants' responses yielded networks with low levels of overall connectivity, including many isolated clusters of just two items. The Hebrew learners, on the other hand, produced much longer chains of interconnected responses, with fewer isolated clusters. Schur interprets these in terms of language learning background. Citing 'informal feedback' from the Chinese participants and their teachers, she suggests that the Chinese learners' shallow networks may be the result of a rote style of learning which placed little emphasis on communication, whereas the Hebrew learners had learnt English in a much more communicative context. Schur suggests that this may have made them more aware of the multiple meanings of words and different potential links between them (see Section 5.3 for further discussion of language learning background on associative knowledge).

Finally, it is worth reiterating that when researchers create WA-based networks, they are creating networks of WA response data, not accurate models of the human mind: the network model remains a metaphor. This is evident in the way that different methodological choices, such as the use of single vs. multiple response tasks, lead to different network properties. Expanding on this, Wilks & Meara (2007) have questioned whether WA tests

provide the ‘direct access’ to the lexicon that they are often assumed to. Much WA research, both L1 and L2, appears to define two words as being ‘associated’ simply because one yields the other in a WAT. Wilks and Meara suggest that this is not necessarily the case, since WA responses tend to vary depending on respondent strategies (2007; Riegel et al. 1967) and task demands (see for example Suzuki-Parker & Higginbotham, 2019, on differences between oral and written WA responses). Similarly, De Deyne, Verheyen, & Storms (2016) have demonstrated that lexical networks built from corpus data differ in fundamental ways from those built from WA data. Research such as this provides a reminder that no single method of lexical network construction is likely to perfectly replicate the contents of the human mental lexicon.

5.2. The influence of lexical variables on word association

This section explores the numerous ways in which the semantic and distributional properties of words influence the WA task. We here take the view that the word association task is a form of psycholingustic elicitation, and as such reflects properties of cue words in ways which may not be true of other psycholinguistic processes (Mollin, 2009, Nordquist, 2009). The network studies described in the previous section have been important to the endeavour of exploring these properties, since their holistic view has facilitated a shift from seeing lexical variables as influencing only local connections between words, to an approach which sees lexical properties as a key determinant of the structure of the associative network. The vast majority of research conducted in this area has been carried out using L1 respondents, and to date very few studies have explored the extent to which L1 findings are applicable to L2. These are discussed at the end of this section; the following review of lexical variable effects is largely informed by L1 literature.

Numerous distributional and semantic properties of words have been shown to influence WA. The most widely researched of these characteristics are described in Table 3.

Table 3 Lexical variables.

Measure	Description
Distributional measures	
Frequency	How often a word occurs in a corpus; typically measured as the number of occurrences per million words of corpus text (fpmw).
Contextual diversity	The number of different documents in a corpus in which a word appears; recently argued to provide a better measure of a word's salience than frequency (Adelman, Brown & Quesada 2006).
Age of acquisition (AoA)	An estimate of the age at which a word is typically first acquired. Generated via group norming procedures (e.g. Bird, Franklin & Howard 2001).
Semantic measures	
Grammatical class (GC)	The class of a word, such as noun, verb, and adjective.
Concreteness	The extent to which a word's meaning can be perceived through the senses. Ranges from highly concrete (e.g. *peacock*) to highly abstract (e.g. *belief*). Generated via group norming procedures (e.g. Brysbaert, Warriner & Kuperman 2014)
Imageability	A measure of how easy it is to generate a mental image of a concept. For example, *table* is more imageable than *justice*. Generated via group norming procedures; highly correlated with concreteness (Bird et al. 2001).
Affective variables	A cluster of variables generally comprising: - valence (how positive a word is felt to be; high valence: *Christmas*; low valence: *torture*) - dominance (the sense of control a respondent feels over a concept; high: *project*; low: *earthquake*) - arousal (the degree of activeness of a concept; high: *tornado*; low: *asleep*) Generated via group norming procedures (e.g. Warriner, Kuperman & Brysbaert 2013)

An entry point into research on these variables is a study by De Deyne & Storms (2008). Their experiment involved the creation of a large network of Dutch L1 WA responses

collected using 1424 cue words. The network was analysed in order to determine the extent to which the lexical properties of its constituent words influenced its structure. Two measurements are particularly useful for picking apart these lexical effects: out-degree and in-degree (see Section 5.1 for definitions). De Deyne & Storms found that lexical variables interact with these measures in four ways:

1. Distributional variables correlate only weakly, or not at all, with out-degree (Frequency r=.14, p<.01; AoA r=.02, p>.05);
2. Distributional variables correlate much more strongly with in-degree (Frequency r=.7; AoA r=-.64, both p<.01);
3. Only one semantic variable, imageability, was used in the study; it did not correlate significantly with out-degree;
4. Imageability did, however, correlate with in-degree (r=.30, p<.01).

De Deyne & Storms' findings offer a holistic perspective against which we can evaluate findings from other studies:

Finding one, above, might imply that the distributional properties of cue words hold only a very modest influence over the generation of responses. Other studies have generally supported this implication. Studies on cue frequency, for example, have revealed only weak effects: a seminal study by de Groot (1989) found slightly slowed responses and greater response heterogeneity for higher frequency cues, but these effects were modest and emerged only when cue frequencies were very widely spaced. Compounding this weak effect are contradictory findings presented in earlier reviews by Brown (1971) and Cramer (1968), which suggested that higher frequency cues sometimes yield slightly *faster* responses. Both Brown and de Groot conclude that cue frequency has only a marginal effect on response patterns. Indeed, Stolz & Tiffany (1972) suggest that WA frequency effects are better viewed as word *familiarity* effects than as pure frequency effects – a suggestion which has been

important to L2 WA researchers (see below). Research into the influence of AoA is slightly more clear: both Brysbaert, Van Wijnendaele & De Deyne (2000) and van Loon-Vervoorn (1989) find that WA responses are produced more quickly when the cue is an early acquired word. Finally, the view that distributional data holds little influence over response patterns is supported by Van Rensbergen, Storms & De Deyne (2015), who found that a cue's distributional properties were very poor predictors of the same properties of their responses (AoA R^2=.04; Contextual diversity R^2=.01; Frequency R^2=.01).

The second finding emerging from De Deyne & Storms' research suggests that frequent and early acquired words are much more likely than others to assume the role of hubs in the associative network, since they appear to possess a higher number of outgoing links than less frequent or later acquired words do. Although few studies have attempted to replicate these findings, they are nevertheless given conceptual validity by the theory of preferential attachment, which suggests that early acquired words develop high centrality because they provide an anchor for new knowledge (Steyvers & Tenenbaum 2005, and see Section 5.1). This is an area with important implications for L2 research. If words learned early in life help to anchor lexical knowledge in an L1, then it is useful to ask whether these same words play a similar role in L2 acquisition; more generally, what are the properties of words central to L2 lexical networks? The contrasting influence of distributional variables when measured as *cue* vs *response* properties also point to the importance of large-scale network research: such findings are simply invisible to microscopic research into local connections between isolated words.

Thirdly, De Deyne & Storms' research suggests that a cue word's imageability (and perhaps, by extension, other semantic variables) holds no significant influence over response generation. This implication is not, however, supported by WA studies using a wider range of methods. An initial indication of the importance of cue semantics to WA response patterns is

given by Van Rensbergen, Storms & De Deyne (2015). They showed that while a cue's distributional properties were poor predictors of the same properties of responses, the cue's semantic features mirrored response properties to a much greater extent (Concreteness R^2=.20; Valence R^2=.31; Arousal R^2= .17; Dominance R^2=.15; all p<.001). Further examples of the importance of specific semantic properties to response generation include:

- Concrete cues have been shown to be responded to more quickly (Brown 1971; de Groot 1989; van Hell & de Groot 1998), receive fewer blank responses (de Groot 1989; Bøyum 2016), yield more homogeneous distributions (Brown 1971; de Groot 1989), and have higher response availability (de Groot 1989) than less concrete ones.
- Van Rensbergen et. al. (2016), taking a mesoscopic approach, have shown that WA networks can be used to generate norms for valence, dominance, and arousal which correlate to a high degree with human judgements of affective strength; reviews by Cramer (1968) and Brown (1971) offer further examples of the influence of affective variables.

The last of the findings in the list above implies that words with strong semantic profiles (e.g. high imageability or strongly affective properties) may be more central to the lexicon than more semantically opaque words (De Deyne & Storms 2008). Van Rensbergen et. al.'s (2016) work on generating affective norms from WA data appears to support this implication. However, as with the distributional data described above, the lack of research on outgoing associative connections to date means that it is too early to draw any conclusions.

The L1 findings presented above suggest that WA may offer a promising method for exploring the properties of second language lexical structure, for example by identifying the properties of words which are central to the L2 lexicon. It is critically important, however, not to assume that L1 findings will transfer straightforwardly to L2. Second languages are typically less completely acquired than first languages, and this is reflected in findings

pertaining to frequency effects in L2 WA. Several studies have found that low frequency cues lead to less consistent responses (Meara 2009), more blank responses (Higginbotham 2010), and lower response availability (Zareva 2011) than high frequency words, suggesting that WA in a second language is more sensitive to frequency effects than L1 association. This has led some researchers to adopt Stolz & Tiffany's suggestion that these findings may be better understood as word familiarity effects. For example, Wolter (2001) demonstrated that both first and second language respondents produce increasing numbers of form-based responses as word familiarity declines, while Riegel and Zivian (1972) found a similar pattern among trilingual participants (see also Wilks 2009). Results such as these led Zareva (2011) to call for dedicated studies disambiguating the role of cue frequency and familiarity in L2 WA.

Another justification for rejecting assumptions of L1 and L2 associative similarity can be derived from the theory of preferential attachment (see Section 5.1). If early-acquired vocabulary knowledge provides an anchor for later learning in an L1, it follows that newly learned L2 items may also preferentially attach to L1 words or concepts. At least two studies have hinted that this might be the case, and that lexical variables such as concreteness might in fact be one determinant of the extent of such linguistic and conceptual entanglement (see Section 5.3). Firstly, Kolers (1963) found that bilinguals producing responses in both of their languages generated a higher number of translation equivalents when the cue in question was a concrete word than when it was more abstract. Secondly, research by van Hell and de Groot (1998) found that concrete cues were more likely than abstract ones to result in participants producing the same response in both L1 and L2 tasks. Responses were also produced more quickly to concrete cues in both L1 and L2. Van Hell & de Groot additionally imply that the existence of (L1) conceptual knowledge might be a complicating factor in the interpretation of L2 WA research.

The comparison of L1 and L2 WA studies in this section has exposed a critical methodological requirement for second language WA work: approaches should be developed which allow the unique role of each lexical variable to become clear. Several studies investigating the role of grammatical class (GC) in L2 WA stand in evidence of this. In L1 studies, numerous effects of this variable have been identified in both micro- and macroscopic studies. In general, these findings suggest that nouns are more central to associative networks than words of other classes, since they tend to be the most common type of response to noun, verb, and adjective cues, and dominate lists of hubs in associative networks (Deese 1962a; Entwisle 1966; Cramer 1968; De Deyne & Storms 2008, 2015). Methodological issues have obscured such findings in L2 research, however. For example, Nissen & Henriksen (2006) found that the pattern of influence from cue GC in L2 WA mirrors that found in L1. However, they also acknowledged that their experimental design did not control for the effects of important variables such as concreteness and age of acquisition, leaving some question marks over their findings. Similarly, Zareva (2010, 2011) reported an interaction between participant proficiency, cue GC, and cue frequency, which resulted in significant differences in syntagmatic and paradigmatic responses, but was unable to identify the specific locus of this interaction, partly because of the lack of statistical power resulting from her use of only 12 cues from each GC. These points underline the need for methodological clarity in future L2 WA studies.

Perhaps the most important methodological concern facing research into the effect of lexical variables on L2 associative structure is the lack of large-scale L2 network-based WA research. While some researchers have attempted to apply a network perspective to their research, they have tended to use simulation methods rather than empirical data (e.g. Meara 2006; 2007; though see Schur, 2007, and Zareva, 2010, for small-scale network-oriented behavioural studies). What this means in practice is that while L1 research has provided good

evidence that various properties of words strongly influence the structure of the lexicon, for example by providing anchors for the integration of new knowledge, none of the structural properties of L1 associative networks described in the first half of this section can yet be asserted in an L2.

5.3. Mapping word association findings onto psychological constructs

Scholars have long been interested in the psycholinguistic systems and processes which underlie the formation and production of word associations, as documented in Section 2. The classical assumption regarding the formation of associations was that words become associated in the mind through textual contiguity – that is, when they co-occur in texts (see Deese 1966, and Warren 1916, for historical overviews). This belief has led some researchers to argue that the generation of word associations also involves recall of these co-occurring words. For example, Wettler, Rapp, & Sedlmeier (2005: 116) have shown that a corpus-derived model of WA can predict a number of primary human WA responses, and generally behaves in a manner similar to human WA respondents. The authors conclude that 'the behaviour of participants in the free association task can be explained by associative learning of the contiguities between words'.

This is not the view of all scholars, however. Others point to the weakness of correlations between corpus co-occurrence and WA response strengths (e.g. Mollin 2009; Kang 2018), and warn against assuming that the processes underlying the *learning* of association are necessarily also the ones responsible for the retention and generation of words in WATs (Hutchison 2003; McRae, Khalkhali, & Hare 2012). These researchers argue that while textual contiguity may be, in the words of Kang (2018: 110), a 'starting point' for the *development* of associative knowledge, the lexical information drawn on during WA response

generation is in fact semantic in nature (Guida & Lenci 2007; Mollin 2009; McRae et al. 2012; De Deyne & Storms 2015; Thwaites 2019). According to this view, contiguities between words provide input which is then acted upon by the mind. WA responses therefore reflect not the contiguities themselves, but the mental system which results from this semantically-driven cognitive activity (cf. Deese 1966). Correspondences between corpus and WA data such as those presented by Wettler et. al. (2005) are explained as reflecting the co-occurrence of semantically related words in text. Evidence in support of this view includes the following:

- Priming effects have been reported between words which are semantically related but do not co-occur: no reliable priming effects have been demonstrated for words which co-occur but are not semantically related (Hutchison 2003);
- While several studies have demonstrated that non-semantic associations can be learned under experimental conditions, the process is time-consuming, and the resulting associations are not generalised to new tasks (Dagenbach, Horst, & Carr 1990; Schrijnemakers & Raaijmakers 1997);
- Associations presented in sentential contexts are learnt more easily than decontextualized ones (Prior & Bentin 2003); the process of integrative semantic processing of sentences can account for the retention of associations (Prior & Bentin 2008);
- Most studies comparing WA responses to corpus data reveal significant differences between the two types of data (Mollin 2009; Kang 2018); word associations tend to reflect prototypical interpretations of words (e.g. *erupt-volcano*), while figurative uses (e.g. *erupt-violence*), often found in corpora, are uncommon in WA (Thwaites 2019);
- Responses to concrete noun cues in WA correspond to a high degree (72.5%) with defining semantic features of their cue (Vivas et al. 2018);

- Semantic aspects of cue words (e.g. verbs pertaining to weather, bodily processes, or motion) influence proportions of responses of different types (e.g. synonyms, thematic roles; Guida & Lenci 2007).

This evidence suggests that the process of generating WA responses, at least in a first language, is largely a semantic one.

The neural basis of this semantic knowledge has attracted much research, both in WA and the wider psycholinguistic research community. Two broad viewpoints have been put forward. The first is the AMODAL position. Researchers such as Fodor (1975; 1989) have suggested that words exist in the mind within a single dedicated neural system which organises them in relation to one another. The system works by converting linguistic experience, such as the books we read or the conversations we have (as well as, in some amodal models, aspects of our perceptual experience such as sensory input and motor activity), into a network of semantic symbols unrelated to perceptual brain states. Expressed somewhat crudely, amodal systems posit that we know what words mean because we are aware of their similarities and differences to other words.

The opposing, MODALITY-SPECIFIC, view, which encompasses theories of embodied cognition, is that words are not stored in a single brain system, but are distributed across the brain together with the motor, perceptual, and introspective experiences to which they correspond. Processing words such as *grasp*, according to these models, will therefore activate those neural systems which underlie the physical act of grasping (e.g. Wilson 2002; Bergen & Chang 2004; Barsalou 2008; see also the *Distributed Lexicon* model in Wray 2002). Proponents of this view have suggested that these models provide *a priori* explanations of numerous linguistic phenomena which are less satisfactorily explained by amodal models (see Martin 2007 for a review; also Barsalou 2008).

Evidence in support of both of these viewpoints has been provided in WA studies. For example, the viability of the amodal position as an explanation of L1 associative network properties is suggested by Gruenenfelder et. al. (2016), who found that these properties can be replicated using only corpus data (i.e. without the need for additional perceptual information). On the other hand, several studies have lent support to one particular modality-specific model: Barsalou's Language and Situated Simulation model (LASS; Barsalou et al. 2008). LASS posits that discrete neural systems handle linguistic and conceptual knowledge; linguistic events (including WA) activate both of these systems. Importantly, however, the former is accessed more quickly than the latter. This leads to a number of predictions. Firstly, it should be possible to identify two distinct patterns of brain activity during WA, corresponding to linguistic and conceptual processing respectively, since separate neural systems are posited to handle these. Secondly, there should be a different time course for the responses originating from the two different systems, with conceptual responses produced more slowly. Both of these predictions have been supported experimentally (De Deyne & Storms 2008; Simmons et al. 2008; Santos et al. 2011), though in some cases methodological issues such as the use of novel categorization schemes mean that these findings are in need of independent replication.

A somewhat intermediary position between modal and amodal views is offered by 'hub-and-spoke' models (Jackson et al. 2015), which propose an integrative, dynamic semantic system (the 'hub') which collects information from linguistic, introspective, and perceptual sources within a single system (and is therefore somewhat similar to the system proposed in amodal models), while retaining links (the 'spokes') to modality-specific systems which contain more specific conceptual information. Such hybrid systems, which do not insist upon hard divisions between linguistic and perceptual processing, offer a potentially exciting alternative to the models described above, but are yet to be explored in a WA context

(Andrews, Vigliocco, & Vinson 2009; Andrews, Frank, & Vigliocco 2014; Fernandino et al. 2016).

Several models of semantic knowledge have the potential to contribute to debates regarding the AUTOMATICITY of lexical processing in WA. LASS, for example, may imply that linguistic WA responses are somewhat more automatic than conceptual ones, since the latter might be slowed by conscious generation of images and memories.

Playfoot et al. (2016) looked more closely at the concept of automaticity, investigating whether WA responses are in fact the first word which comes to mind, as is generally assumed. Across two experiments, the authors found working memory capacity effects on both a standard free association task and a novel creative association task, in which participants were asked to generate associations which they felt would be unique. Individuals with higher working memory capacity produced more stereotypical responses on the free task, and more unique responses on the creative task. Since this interaction of task demands and working memory potentially suggests strategic responding (i.e. selection of a response which was not the first which came to mind), the authors designed a further experiment in which working memory use was inhibited. They found that the proportion of stereotypical responses produced per participant increased in this time-constrained condition. This appears to suggest that the stereotypical responses generated in the free response task are in fact the first word that comes to participants' minds. Playfoot et al.'s study therefore suggests some level of automaticity in WA response generation.

Studies on psycholinguistic aspects of *second* language WA have tended to focus on the influence of L1 lexical knowledge on L2 association. Some studies have attempted to link various aspects of associative response patterns to L2 teaching methods. An early example of this type of research is a study by Riegel et. al. (1967), who compared responses from bilingual (L1 English, L2 Spanish; L1 Spanish, L2 English) participants on two tasks: a

traditional WA task, and a constrained task in which respondents were asked to produce specific types of association, such as words which shared semantic features with the cue. The authors note the different linguistic background of the two sets of participants: the Spanish L2 group had learned in a classroom environment, while the L2 English participants (who reported slightly lower proficiency) were living in the US and using English communicatively. They found that the L2 English participants gave fewer blank responses and showed greater response homogeneity than the Spanish L2 group. However, their responses also showed less overlap with English L1 responses on the constrained WA task than did the L2 Spanish group. According to the authors, this implied that while the L2 English group had a more fluent grasp of word meanings, the Spanish L2 group possessed more detailed conceptual representations of word meanings. Riegel et. al. suggest that this might be due to the different L2 acquisition contexts, with an immersion context leading to greater lexical fluency and a classroom one to greater lexical precision.

While this line of enquiry has not yet been extensively pursued, it is worth recalling, firstly, the findings of Schur (2007; see Section 5.1), who suggested in the light of L2 associative network research that learning English in a communicative classroom may have afforded her Israeli participants more interconnected lexical networks than the Chinese participants who had studied in classrooms dominated by rote learning of vocabulary; and secondly those of Håkansson & Norrby (2010), who found that language learning context influenced the number of clang responses produced by L2 WA participants (Section 4). Both of these findings additionally resonate with the predictions of usage-based theories of language (see Section 3.2) that the nature of linguistic experience shapes the structure of linguistic knowledge.

Riegel et. al.'s (1967) study makes an important assumption – that L2 WA reflects the degree to which L2 conceptual representations have been acquired. This view implies that

language learning necessarily involves the development of new conceptual representations for words, corresponding to those of speakers of the target language. A contrasting view is that language learners do not develop new representations, but instead simply apply their pre-existing L1 representations to new L2 words. Several studies have argued for the latter, suggesting that three cue properties – concreteness, cognate status, and grammatical class – might mediate the extent to which new L2 representations are developed (Kolers, 1963; Taylor, 1976).

In an important examination of this issue, van Hell & de Groot (1998) found that bilingual (Dutch L1, English L2) participants were most likely to give the same response in both languages to concrete cognate nouns; abstract cognates, concrete non-cognates, and abstract non-cognates respectively yielded declining levels of response repetition, as did non-noun cues. The authors interpreted these findings as supporting a distributed, network model of conceptual representation, according to which the 'conceptual units' which make up word meanings are shared between languages only where they overlap. Since there is greater between-language overlap for concrete than for abstract concepts, for cognate than for non-cognate words, and for nouns than for other grammatical classes, it is considered that conceptual representations for these words (concrete, cognate, nouns) are more likely to be shared between languages. From a language teaching perspective, this may also suggest that time spent helping learners to develop an awareness of conceptual differences between cognates may be beneficial.

A different approach to studying the relationship between first and second language lexical access was taken by Fitzpatrick & Izura (2011). Their study had two aims, firstly to test whether responses of different categorisations (meaning-, position-, or form-based, or a combination of two of these) differed in response speed (in L1 and/or L2); and secondly, to test (via a lexical decision task) whether L1 translation equivalents had been accessed during

L2 response generation. The study also tested whether L2 vocabulary knowledge (as a proxy for proficiency) influenced these measurements. The participants were 24 bilinguals (L1 Spanish, L2 English) living in the UK.

A number of findings emerged from the study. Firstly, L1 response times (RT) were faster than L2 response times for the corresponding categories. Secondly, the fastest responses (irrespective of response language) were words which were both meaning- and position-based (i.e. dual category responses such as *pen>paper*); responses of non-equivalent meaning to the cue (e.g. *party>celebrate*; *accountant>numbers*) were slower than other response types. Thirdly, L2 participants with higher vocabulary test scores responded more quickly than those with lower tests scores, except in the case of the aforementioned meaning- and position-based dual category responses. Finally, an L1 priming effect was discovered: lexical decision RTs were faster for L1 translation equivalents of L2 cues than for L1 filler words. The priming effect appeared to be influenced by proficiency level, since it was only statistically significant for participants who scored below average on the test of vocabulary knowledge. Fitzpatrick & Izura's results are therefore supportive of accounts of bilingual lexical processing which view L2 representations as parasitic, at least in the early stages of learning, upon L1 representations. Examples of this type of theory include the Revised Hierarchical Model (Kroll & Stewart 1994) and MacWhinney's (2005) Unified Model of Language Acquisition. The study also suggests the value both of using response time measures in WA, and of testing WA findings using other psycholinguistic measures, such as lexical decision tasks.

While the L2 studies above are too few in number to offer a decisive view of the psycholinguistics of L2 WA, they do suggest significant potential for future research. The influence of learning environments (e.g. classroom methodologies and periods of immersion),

and the role of L1 knowledge in L2 development may be particularly worthy of further investigation.

6. Conclusion

This paper represents the first comprehensive scrutiny and evaluation of WA research from different disciplinary domains and research eras. Juxtaposing findings from conceptually different approaches has afforded novel insights and perspectives, and the assertions and suggested aims we set out below emerge from this.

Beginning our paper with a historical perspective enabled us to draw attention to the circularity of some lines of investigation; there are significant overlaps in terms of research aims, methods and (inconclusive) findings between work conducted 100+ years ago and that published in the last few decades. This is particularly the case in connection with the stereotypy and categorization studies reported in Sections 3 and 4; our long view of contributions to that 'microscopic' research strand finds little cause for optimism that a methodological breakthrough will yield meaningful findings in connection with language proficiency or other variables hitherto explored, at least in investigations of group data from the populations typically targeted.

This is not to say that further small incremental improvements cannot be made to research designs, but the multiple influences on response behaviour at an individual word level, some of which are nigh impossible to discern, make too much white noise for clear relationships with discrete variables to be identified. Responses are a product of interaction between the experience/characteristics of an individual, the properties of cue words, and the features of specific languages. Careful selection of participant groups and/or cue words, longitudinal within-subject studies, and comparative studies between languages can manage these variables somewhat. However, the instability of the relative weights of these influences

means that it is not feasible to prescribe target responses, for example as benchmarks of proficiency, or to reliably predict WA behaviour from specified variables (or vice versa). A further warning is sounded by Schmitt, Nation & Kremmel (2019) who, noting that test revisions should not be uncritically welcomed, call for revised versions to be subjected to the same validation criteria as new tests. The pattern of iterative refinements to test formats that we have seen in relation to stereotype and WAF measures in particular are cause for concern in this regard.

Having said this, we note that the research conducted in that 'microscopic' strand has tended to take WA data from groups of typical language users, and to seek correlations with group membership (L1/L2 user, high/low proficiency, old/young, etc.). WA data from groups of *atypical* language users, especially data controlling the lexical variables in Table 3 above, might have capacity to cut through some of that white noise. For example, we might predict that WA responses from visually impaired participants, especially if they have been blind from birth, will not be affected by imageability (see Metcalfe 2019); WA responses from people with dementia, whose language has begun to attrite, might be affected in specific ways by the age of acquisition or the frequency variables, and so on. Confounding effects of individual differences can be eliminated through longitudinal, within-subject studies, and examining changes in WA behaviour over time can help us understand effects of increasing proficiency, or extreme language-related events (e.g. study abroad periods) on an individual's lexical networks.

Despite the frustrating dearth of consistent outcomes from studies using WA to investigate lexical knowledge and retrieval, this paper has identified some promising pockets of research. We propose that two lines of investigation in particular have significant capacity to drive different ways of thinking about WA, and to generate new research agendas. First, network approaches to L1 WA research have enjoyed considerable success, and have begun

to reveal some of the foundational principles and structural properties of associative networks, as well as the lexical determinants of this structure. We propose that similar research with L2 participants might therefore shed light on key aspects of their lexical development, which in turn has potential to reveal new insights into i) the design of lexical syllabi or teaching materials through attention to centrality and hub words; ii) the influence of language teaching methodology on network properties; and/or iii) the relationship between the properties themselves and receptive or productive lexical fluency. A further benefit of such research is that it can be designed to investigate both individual and group-level networks. As such, although some researchers have highlighted the difficulty of collecting L2 network data (e.g. Wilks & Meara 2007; Meara 2009; Wilks 2009), we believe that renewed efforts at modelling L2 lexical networks using WA data are likely to be worth the effort.

Second, our review of a broad sweep of WA-related literature has yielded surprisingly few attempts to model WA response generation from a theoretical point of view, either in first or additional languages. Two recent studies (Bøyum 2016; Thwaites 2019) have begun to address this by viewing WA from a usage-based perspective. Thwaites (2019), for example, presents evidence suggesting that the orderliness of a word's collocational distribution may, along with semantic processing of lexical items, be a significant determinant of associative response type and distribution (see also Hahn & Sivley 2011; Kang 2018). This suggests, at the very least, that a usage-based approach to WA which views the structure of lexical knowledge as being the result both of the mind's sensitivity to probabilistic aspects of language and its capacity for sorting and categorising input, may provide fertile ground upon which to cultivate further research. The intersection of corpus-based networks and psycholinguistic networks might be usefully explored in this regard. Such data has capacity to afford comparative semantic analysis (of similarity between items, for example), and might be extended to natural language processing models (NLP).

This paper has highlighted areas that we consider unlikely to be fruitful lines of further enquiry, and has exposed questions about aspects of WA research that have not yet been fully explored, but have potential to yield findings that break through current impasses in this research domain. We conclude with a list of priority research questions; in addressing these we believe that word association researchers can develop robust new theoretical paradigms, use new findings to creatively extend learning and teaching practices, and affirm the contribution of WA research to our understanding of the workings of the mental lexicon.

7. Priority research questions arising from this paper

Language Learning and Teaching

- Is there a causal relationship between the methods by which learners study vocabulary (e.g. rote learning vs. communicative methods) and their word association (network) properties?
- What is the relationship between small world network properties and lexical fluency, and can this inform L2 vocabulary learning and teaching?
- How might the differences in the quality and quantity of linguistic input experienced in first and (classroom-based) second language acquisition influence the responses given in WA tasks? Does data from these groups support a usage-based model of processing?
- How do learning environments (e.g. classroom methodologies and periods of immersion) influence the structure of L2 lexical knowledge; and how does that lexical structure influence receptive and productive language use?
- What is the role of centrality (preferential hub words) in L2 acquisition, and can its power be harnessed in order to escalate vocabulary uptake?

Second Language Acquisition

- (How) are second language associative networks different from first language networks?
- Which words are most central to L2 associative networks? Are they the same (and do they have the same general properties) as those which are central to L1 networks?
- If early acquired words serve as anchors for later lexical knowledge in an L1, what anchors L2 lexical knowledge?

Language Specificity

- To what extent are existing word association findings replicable, within and between languages?
- Given that English is the language context of the majority of WA research, is it possible to distinguish findings specific to the linguistic features of English, from non-language-specific findings?

Beyond language learning and SLA

- How might networks of WA data be of assistance to NLP researchers, whose attempts to facilitate interaction between humans and computers continues to be based largely on corpus data?
- Can a model be created that accounts for WA responses as an interaction between corpus derived data (that which an individual is exposed to) and psycholinguistic processes in such a way that variations in either can be identified in WA responses?

References

Adelman, J. S., G. D. A. Brown, & J. F. Quesada (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science* 17.9, 814–823.

Aitchison, J. (2012). *Words in the mind: an introduction to the mental lexicon* (4th edn.). Chichester: John Wiley & Sons.

Albrechtsen, D., K. Haastrup, & B. Henriksen (2008). *Vocabulary and writing in a first and second language: Processes and development*. Basingstoke: Palgrave Macmillan.

Andrews, M., S. Frank, & G. Vigliocco (2014). Reconciling embodied and distributional accounts of meaning in language. *Topics in Cognitive Science* 6.3, 359–370.

Andrews, M., G. Vigliocco, & D. P. Vinson (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review* 116.3, 463–498.

Andrews, M., D. P. Vinson, & G. Vigliocco (2008). Inferring a probabilistic model of semantic memory from word association norms. In *Proceedings of the 30th annual conference of the Cognitive Science Society*, 1941–1946.

Bally, C. & A. Sechehaye (1966). *Course in general linguistics Ferdinand de Saussure*. McGraw-Hill Book Company.

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology* 59, 617–645.

Barsalou, L. W., A. Santos, W. K. Simmons, & C. D. Wilson (2008). Language and simulation in conceptual processing. In M. De Vega & A. M. Glenberg (eds.). *Symbols, embodiment, and meaning*. Oxford: Oxford University Press, 245–283.

Batty, A. (2012). Identifying dimensions of vocabulary knowledge in the word associates test. *Vocabulary Learning and Instruction* 1, 70–77.

Beckage, N. M., L. B. Smith, & T. Hills (2010). Semantic network connectivity is related to

vocabulary growth rate in children. *Proceedings of the 32nd Annual Meeting of the Cognitive Science Society*, 2769–2774.

Bergen, B. K. & N. Chang (2004). Embodied construction grammar in simulation-based language understanding. In J. Östman & M. Fried (eds.) *Construction grammars: cognitive grounding and theoretical extensions*. Amsterdam: John Benjamins, 147-189.

Bird, H., S. Franklin, & D. Howard (2001). Age of acquisition and imageability ratings for a large set of words, including verbs and function words. *Behavior Research Methods, Instruments, & Computers* 33.1, 73–79.

Bøyum, M. (2016). A new word association test for Norwegian. MA Dissertation, University of Oslo.

Brown, R. & J. Berko (1960). Word association and the acquisition of grammar. *Child Development* 31.1, 1–14.

Brown, W. P. (1971). A retrospective study of stimulus variables in word association. *Journal of Verbal Learning and Verbal Behavior* 10.4, 355–366.

Brysbaert, M., A. B. Warriner, & V. Kuperman (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods* 46.3, 904–911.

Brysbaert, M., I. Van Wijnendaele, & S. De Deyne (2000). Age-of-acquisition effects in semantic processing tasks. *Acta Psychologica* 104.2, 215–226.

Bybee, J. & C. Beckner (2010). Usage-based theory. In B. Heine & H. Narrog (eds.). *Oxford handbook of linguistic analysis*. Oxford: Oxford University Press, 827–856.

Cramer, P. (1968). *Word association*. New York, NY: Academic Press.

Cremer, M., D. Dingshoff, M. de Beer, & R. Schoonen (2011). Do word associations assess word knowledge? A comparison of L1 and L2, child and adult word associations. *International Journal of Bilingualism* 15.2, 187–204.

Cremer, M. & R. Schoonen (2013). The role of accessibility of semantic word knowledge in

monolingual and bilingual fifth-grade reading. *Applied Psycholinguistics* 34.6, 1195–1217.

Dagenbach, D., S. Horst, & T. H. Carr (1990). Adding new information to semantic memory: how much learning is enough to produce automatic priming? *Journal of Experimental Psychology: Learning, Memory, and Cognition* 16.4, 581–591.

Deese, J. (1962). On the structure of associative meaning. *Psychological Review* 69.3, 161–175.

Deese, J. (1966). *The structure of associations in language and thought*. Baltimore: John Hopkins University Press.

De Deyne, S., D. J. Navarro, A. Perfors, M. Brysbaert, & G. Storms (2018). The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior Research Methods* 51.3, 987-1006.

De Deyne, S., D. J. Navarro, & G. Storms (2012). Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. *Behavior Research Methods* 45.2, 480–498.

De Deyne, S. & G. Storms (2008). Word associations: network and semantic properties. *Behavior Research Methods* 40.1, 213–231.

De Deyne, S. & G. Storms (2015). Word Associations. In J. R. Taylor (ed.). *The Oxford handbook of the word*. Oxford: Oxford University Press, 465–480.

De Deyne, S., S. Verheyen, & G. Storms (2016). Structure and organization of the mental lexicon: A network approach derived from syntactic dependency relations and word associations. In A. Mehler, A. Lücking, S. Banisch, P. Blanchard, B. Job (eds.) *Towards a theoretical framework for analyzing complex linguistic networks. Understanding complex systems.* Berlin, Heidelberg: Springer, 47-79.

de Groot, A. M. B. (1989). Representational aspects of word imageability and word

frequency as assessed through word association. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 15.5, 824–845.

Dronjic, V. & R. Helms-Park (2014). Fixed-choice word-association tasks as second-language lexical tests: What native-speaker performance reveals about their potential weaknesses. *Applied Psycholinguistics* 35.1, 193–221.

Duñabeitia, J. A., A. Avilés, & M. Carreiras (2008). NoA's ark: Influence of the number of associates in visual word recognition. *Psychonomic Bulletin & Review* 15.6, 1072–1077.

Ellis, N. C., U. Römer, & M. B. O'Donnell (2016). *Usage-based approaches to language acquisition and processing: cognitive and corpus investigations of construction grammar*. Chichester: John Wiley & Sons.

Entwisle, D. R. (1966). Form class and children's word associations. *Journal of Verbal Learning and Verbal Behaviour* 5, 558–565.

Entwisle, D. R., D. F. Forsyth, & R. Muuss (1964). The syntactic-paradigmatic shift in children's word associations. *Journal of Verbal Learning and Verbal Behavior* 3, 19–29.

Ervin, S. M. (1961). Changes with age in the verbal determinants of word-association. *The American Journal of Psychology* 74.3, 361–372.

Esper, E. A. (1918). A contribution to the experimental study of analogy. *Psychological Review* 25.6, 468–487.

Fernandino, L., J. R. Binder, R. H. Desai, S. L. Pendl, C. J. Humphries, W. L. Gross, L. L. Conant, & M. S. Seidenberg (2016). Concept representation reflects multimodal abstraction: A framework for embodied semantics. *Cerebral Cortex* 26.5, 2018-2035.

Fitzpatrick, T. (2006). Habits and Rabbits: Word associations and the L2 lexicon. In S. H. Foster-Cohen, M. Medved Krajnovic, & J. Mihaljevic Djigunovic (eds.). *EUROSLA Yearbook 6*, 121–45.

Fitzpatrick, T. (2007). Word association patterns: unpacking the assumptions. *International*

Journal of Applied Linguistics 17.3, 319–31.

Fitzpatrick, T. (2009). Word association profiles in a first and second language: Puzzles and problems. In T. Fitzpatrick & A. Barfield (eds.). *Lexical Processing in Second Language Learners*, 38–52.

Fitzpatrick, T. (2012). Tracking the changes: vocabulary acquisition in the study abroad context. *The Language Learning Journal* 40.1, 81–98.

Fitzpatrick, T. & C. Izura (2011). Word Association in L1 and L2: An exploratory study of response types, response times, and interlingual mediation. *Studies in Second Language Acquisition* 33.3, 373–398.

Fitzpatrick, T. & I. Munby (2014). Word associations and the L2 lexicon. In J. Milton & T. Fitzpatrick (eds.). *Dimensions of vocabulary knowledge.* Basingstoke: Palgrave McMillan, 92–105.

Fitzpatrick, T., D. Playfoot, A. Wray, & M. J. Wright (2015). Establishing the reliability of word association data for investigating individual and group differences. *Applied Linguistics* 36.1, 23–50.

Fodor, J. A. (1975). *The language of thought*. Cambridge, MA.: Harvard University Press.

Fodor, J. A. (1989). *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA.: Harvard University Press.

Francis, H. (1972). Toward an explanation of the syntagmatic-paradigmatic shift. *Child Development* 43.3, 949–958.

Galton, F. (1879). Psychometric experiments. *Brain* 2.2, 149–162.

Greidanus, T. & L. Nienhuis (2001). Testing the quality of word knowledge in a second language by means of word associations: types of distractors and types of associations. *The Modern Language Journal* 85.4, 567–577.

Griffiths, T. L., M. Steyvers, & A. Firl (2007). Google and the mind. *Psychological Science*

18.12, 1069–1076.

Gruenenfelder, T. M., G. Recchia, T. Rubin, & M. N. Jones (2016). Graph-theoretic properties of networks based on word association norms: implications for models of lexical semantic memory. *Cognitive Science* 40.6, 1460–1495.

Guida, A. & A. Lenci (2007). Semantic properties of word associations to Italian verbs. *Italian Journal of Linguistics* 19.2, 293–326.

Hahn, L. W. & R. M. Sivley (2011). Entropy, semantic relatedness and proximity. *Behavior Research Methods* 43.3, 746–760.

Håkansson, G. & C. E. Norrby (2010). Environmental influence on language acquisition: Comparing second and foreign language acquisition of Swedish. *Language Learning* 60.3, 628–650.Henriksen, B. (2008). Declarative lexical knowledge. In D. Albrechtsen, K. Haastrup, & B. Henriksen (eds.). *Vocabulary and writing in a first and second language*. New York, NY: Springer, 22–66.

Higginbotham, G. M. (2010). Individual learner profiles from word association tests: The effect of word frequency. *System* 38.3, 379–390.

Hoey, M. (2005). *Lexical priming: a new theory of words and language*. London: Routledge.

Horiba, Y. (2012). Word knowledge and its relation to text comprehension: A comparative study of Chinese-and Korean-speaking L2 learners and L1 speakers of Japanese. *The Modern Language Journal* 96.1, 108-121.

Hui, L. (2011). An investigation into the L2 mental lexicon of Chinese English learners by means of word association. *Chinese Journal of Applied Linguistics* 34.1, 62–76.

Hutchison, K. A. (2003). Is semantic priming due to association strength or feature overlap? A microanalytic review. *Psychonomic Bulletin & Review* 10.4, 785–813.

Jackson, R. L., P. Hoffman, G. Pobric, & M. A. Lambon Ralph (2015). The nature and neural correlates of semantic association versus conceptual similarity. *Cerebral Cortex* 25.11,

4319–4333.

Jiang, N. & J. Zhang (2019). Form prominence in the L2 lexicon: Further evidence from word association. *Second Language Research* 1–22. https://doi.org/10.1177/0267658319827053

Jung, C. G. (1910). The association method. *American Journal of Psychology* 31, 219–69.

Kang, B. (2018). Collocation and word association: Comparing collocation measuring methods. *International Journal of Corpus Linguistics* 23.1, 85–113.

Kent, G. H., & A. J. Rosanoff (1910). A study of association in insanity. *American Journal of Psychiatry* 67.1, 37-96.

Khazaeenezhad, B. & A. Alibabaee (2013). Investigating the role of L2 language proficiency in word association behavior of L2 learners: a case of Iranian EFL learners. *Theory and Practice in Language Studies* 3.1, 108–115.

Kim, M. S. (2013). The mental lexicon of low-proficiency Korean heritage learners. *Heritage Language Journal* 10.1, 17–35.

Kiss, G. R. (1968). Words, associations, and networks. *Journal of Verbal Learning and Verbal Behavior* 7, 707–713.

Kiss, G. R., C. Armstrong, R. Milroy, & J. Piper (1973). An associative thesaurus of English and its computer analysis. In A. J. Aitkin, R. W. Bailey, & N. Hamilton-Smith (eds.). *The computer and literary studies*. Edinburgh: Edinburgh University Press, 153–165.

Kolers, P. A. (1963). Interlingual word associations. *Journal of Verbal Learning and Verbal Behavior* 2.4, 291–300.

Kroll, J. F. & E. Stewart (1994). Category interference in translation and picture naming: evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language* 33.2, 149–174.

Kruse, H., J. Pankhurst, & M. Sharwood Smith (1987). A multiple word association probe in

second language acquisition research. *Studies in Second Language Acquisition* 9.02, 141–154.

Lambert, W. E. (1956). Developmental aspects of second-language acquisition: I. Associational fluency, stimulus provocativeness, and word-order influence. *The Journal of Social Psychology* 43.1, 83-89.

Li, D., X. Zhang & G. Wang (2011). Senior Chinese high school students' awareness of thematic and taxonomic relations in L1 and L2. *Bilingualism: Language and Cognition* 14.4, 444-457.

MacWhinney, B. (2005). A unified model of language acquisition. In A. M. B. de Groot & J. F. Kroll (eds.). *Handbook of bilingualism*. Oxford: Oxford University Press, 49–67.

Martin, A. (2007). The representation of object concepts in the brain. *Annual Review of Psychology* 58.1.

McGee, I. (2009). Adjective-noun collocations in elicited and corpus data: Similarities, differences, and the whys and wherefores. *Corpus Linguistics and Linguistic Theory* 5.1, 79–103.

McRae, K., S. Khalkhali, & M. Hare (2012). Semantic and associative relations in adolescents and young adults: examining a tenuous dichotomy. In V. F. Reyna, S. B. Chapman, M. R. Dougherty, & J. Confrey (eds.). *The adolescent brain: learning, reasoning, and decision making*. Washington, DC: American Psychological Association, 39–66.

Meara, P. (1978). Learners' word associations in French. *Interlanguage Studies Bulletin* 3.2, 192–211.

Meara, P. (1980). Vocabulary acquisition: a neglected aspect of language learning. *Language Teaching and Linquistics Abstracts* 14, 221-246.

Meara, P. (1983). Word associations in a foreign language. *Nottingham Linguistic Circular*

11, 28-38.

Meara, P.M. (1996a). The vocabulary knowledge framework. Unpublished discussion paper available at http://www.lognostics.co.uk/vlibrary/index.htm

Meara, P. (1996b). Dimensions of lexical competence. In G. Brown, K. Malmkjaer, & J. Williams (eds.). *Performance and competence in second language acquisition*. Cambridge: Cambridge University Press.

Meara, P. (2006). Emergent properties of multilingual lexicons. *Applied Linguistics* 27.4, 620–44.

Meara, P. (2007). Simulating word associations in an L2: The effects of structural complexity. *Language Forum* 33.2, 13–31.

Meara, P. (2009). *Connected words: word associations and second language vocabulary acquisition*. Amsterdam: John Benjamins Publishing.

Meara, P. & B. Wolter (2004). V_Links: beyond vocabulary depth. *Angles on the English Speaking World* 4, 85–96.

Metcalfe, J. (2019). *Comparison of the lexical networks of blind and sighted individuals: a preliminary investigation.* MA Dissertation. Swansea University.

Milgram, S. (1967). The small world problem. *Psychology Today* 1.1, 61-67

Mollin, S. (2009). Combining corpus linguistic and psychological data on word co-occurrences: Corpus collocates versus word associations. *Corpus Linguistics and Linguistic Theory* 5.2, 175-200.

Morais, A. S., H. Olsson, & L. J. Schooler (2013). Mapping the structure of semantic memory. *Cognitive Science* 37.1, 125–145.

Mulder, K. & J. H. Hulstijn (2011). Linguistic skills of adult native speakers, as a function of age and level of education. *Applied Linguistics* 32.5, 475–494.

Munby, I. (2011). Development of a multiple response word association test for learners of

English as an L2. Ph.D. dissertation, Swansea University.

Munby, I. (2018). Report on a free continuous word association test (part 3). *HOKUGA* 175, 53–75.

Namei, S. (2004). Bilingual lexical development: A Persian–Swedish word association study. *International Journal of Applied Linguistics* 14.3, 363–388.

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. New York: Newbury House.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nelson, D. L., C. L. McEvoy, & T. A. Schreiber (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods* 36.3, 402-407.

Nelson, K. (1977). The syntagmatic-paradigmatic shift revisited: a review of research and theory. *Psychological Bulletin* 84.1, 93–116.

Nissen, H. B. & B. Henriksen (2006). Word class influence on word association test results. *International Journal of Applied Linguistics* 16.3, 389–408.

Nordquist, D. (2009). Investigating elicited data from a usage-based perspective. *Corpus Linguistics and Linguistic Theory* 5.1, 105–130.

Norrby, C. E. & G. Håkansson (2007). Girl – lass or curl? Word associations in second language learners. *Australian Review of Applied Linguistics* 30.2, 22.1-22.17.

O'Connor, J. (1928). *Born that way*. The Willia. Baltimore.

Orita, M. (2002). Word associations of Japanese EFL learners and native speakers: Shifts in response type distribution and the associative development of individual words. *Review of English Language Education in Japan* 13, 111–120.

Osgood, C. E., T. A. Sebeok, J. W. Gardner, J. B. Carroll, L. D. Newmark, S. M. Ervin, S. Saporta, J. H. Greenberg, D. E. Walker, J. J. Jenkins, K. Wilson, & F. G. Lounsbury

(1954). *Psycholinguistics: A survey of theory and research problems.* Baltimore: Waverly Press.

Palermo, D. S. (1971). Characteristics of word association responses obtained from children in grades one through four. *Developmental Psychology* 5.1, 118–123.

Playfoot, D., T. Balint, V. Pandya, A. Parkes, M. Peters, & S. Richards (2016). Are word association responses really the first words that come to mind? *Applied Linguistics* 39.5, 607-624.

Politzer, R. L. (1978). Paradigmatic and syntagmatic associations of first year French students. In V. Honsa & M. Hardman-de-Bautista (eds.). *Papers in linguistics and child language: Ruth Hirsch Weir memorial volume*. Den Haag: Mouton de Gruyter, 203–210.

Postman, L. J. & G. Keppel (1970). *Norms of word association*. New York NY: Academic Press.

Prior, A. & S. Bentin (2003). Incidental formation of episodic associations: The importance of sentential context. *Memory & Cognition* 31.2, 306–316.

Prior, A. & S. Bentin (2008). Word associations are formed incidentally during sentential semantic integration. *Acta Psychologica* 127.1, 57–71.

Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: an assessment perspective. *Language Learning* 52.3, 513–536.

Qian, D. D. & M. Schedl (2004). Evaluation of an in-depth vocabulary knowledge measure for assessing reading performance. *Language Testing* 21.1, 28–52.

Racine, J. P., G. M. Higginbotham, & I. Munby (2014). Exploring non-native norms: A new direction in word association research. *Vocabulary Education Research Bulletin* 3.2, 28–52.

Randall, M. (1980). Word association behaviour in learners of English as a foreign language.

Polyglot 2.2, B5-D1.

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing* 10.3, 355–371.

Read, J. (1995). Refining the word associates format as a measure of depth of vocabulary knowledge. *New Zealand Studies in Applied Linguistics* 1, 1-17.

Read, J. (1998). Validating a test to measure depth of vocabulary knowledge. In A.J. Kunnan (ed.), *Validation in language assessment*. Mahwah, NJ: Erlbaum, 41-60.

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Richards, J. C. (1976). The role of vocabulary teaching. *TESOL Quarterly* 10.1, 77–89.

Riegel, K. F., R. M. Ramsey, & R. M. Riegel (1967). A comparison of the first and second languages of American and Spanish students. *Journal of Verbal Learning and Verbal Behavior* 6.4, 536–544.

Riegel, K. F. & I. W. M. Zivian (1972). A study of inter- and intralingual associations in English and German. *Language Learning* 22.1, 51–63.

Rosenzweig, M. R. (1961). Comparisons among word-association responses in English, French, German, and Italian. The American Journal of Psychology, 74, 347–360.

Roux, P. W. (2013). Words in the Mind: Exploring the relationship between word association and lexical development. *Polyglossia* 24, 80–91.

Rüke-Dravina, V. (1971). Word associations in monolingual and multilingual individuals. *Linguistics* 9.74, 66–84.

Russell, W. A., & J. J. Jenkins (1954). *The complete Minnesota norms for responses to 100 words from the Kent-Rosanoff Word Association Test*. Technical Report No. 11, University of Minnesota.

Santos, A., S. E. Chaigneau, W. K. Simmons, & L. W. Barsalou (2011). Property generation reflects word association and situated simulation. *Language and Cognition* 3.1, 83–119.

Schmitt, N. (1998a). Quantifying word association responses: what is nativelike? *System* 26.3, 389–401.

Schmitt, N. (1998b). Tracking the incremental acquisition of second language vocabulary: a longitudinal study. *Language Learning* 48.2, 281–317.

Schmitt, N., P. Nation, & B. Kremmel (2019). Moving the field of vocabulary assessment forward: The need for more rigorous test development and validation. *Language Teaching*, 1-12. https://doi.org/10.1017/S0261444819000326

Schmitt, N., J. W. C. Ng, & J. Garras (2011). The word associates format: validation evidence. *Language Testing* 28.1, 105–126.

Schoonen, R. & M. Verhallen (2008). The assessment of deep word knowledge in young first and second language learners. *Language Testing* 25.2, 211–236.

Schrijnemakers, J. M. C. & J. G. W. Raaijmakers (1997). Adding new word associations to semantic memory: Evidence for two interactive learning components. *Acta Psychologica* 96.1–2, 103–132.

Schur, E. (2007). Insights into the structure of L1 and L2 vocabulary networks: intimations of small worlds. In H. Daller, J. Milton, & J. Treffers-Daller (eds.). *Modelling and Assessing Vocabulary Knowledge*. Cambridge: Cambridge University Press, 182–203.

Simmons, W. K., S. B. Hamann, C. L. Harenski, X. P. Hu, & L. W. Barsalou (2008). FMRI evidence for word association and situated simulation in conceptual processing. *Journal of Physiology - Paris* 102.1–3, 106–119.

Singleton, D. M. (1999). *Exploring the second language mental lexicon*. Cambridge: Cambridge University Press.

Söderman, T. (1993). Word associations of foreign language learners and native speakers–different response types and their relevance to lexical development. In B. Hammarberg (ed.). *Problem, Process, Product in Language Learning: Papers from the Stockholm-*

Abo Conference, 21-22 October 1992. Stockholm: Stockholm University, 157–169.

Sommer, R. (1901). *Diagnostik Der Geisteskrankheiten Für Praktische Ärzte Und Studierende [Diagnosis Of Mental Illness For Practical Physicians And Students]*. Urban & Schwarzenberg.

Steyvers, M., R. M. Shiffrin, & D. L. Nelson (2004). Word association spaces for predicting semantic similarity effects in episodic memory. In A. Healy (ed.). *Experimental cognitive psychology and its applications: festschrift in honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer*. American Psychological Association, 237-249.

Steyvers, M. & J. B. Tenenbaum (2005). The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. *Cognitive Science* 29, 41–78.

Stolz, S. & J. Tiffany (1972). The production of "child-like" word associations by adults to unfamiliar adjectives. *Journal of Verbal Learning and Behaviour* 11.1, 38–46.

Strogatz, S. H. (2001). Exploring complex networks. *Nature* 410.6825, 268–276.

Suzuki-Parker, J. & G. Higginbotham (2019). Does method of administration influence word association test responses? *Vocabulary Education Research Bulletin* 8.1, 11-16.

Taylor, I. (1976). Similarity between French and English words - A factor to be considered in bilingual language behaviour? *Journal of Psycholinguistic Research* 5.197, 85–94.

Thumb, A. & K. Marbe (1901). *Experimentelle Untersuchungen Über Die Psychologischen Grundlagen Der Sprachlichen Analogiebildung*. Leipzig: W. Engelmann.

Thwaites, P. (2019). *Lexical and distributional influences on word association response generation.* Ph.D. dissertation, Cardiff University.

van Hell, J. G. & A. M. B. de Groot (1998). Conceptual representation in bilingual memory: Effects of concreteness and cognate status in word association. *Bilingualism: Language and Cognition* 1.3, 193–211.

van Loon-Vervoorn, W. A. (1989). *Eigenschappen van Basiswoorden [Properties of Basic*

Words]. Lisse, The Netherlands: Swets & Zeitlinger.

Van Rensbergen, B., S. De Deyne, & G. Storms (2016). Estimating affective word covariates using word association data. *Behavior Research Methods* 48.4, 1644–1652.

Van Rensbergen, B., G. Storms, & S. De Deyne (2015). Examining assortativity in the mental lexicon: Evidence from word associations. *Psychonomic Bulletin & Review*, 22.6, 1717-1724.

Vivas, L., L. Manoiloff, A. M. Garcia, F. Lizarralde, & J. Vivas (2018). Core semantic links or lexical associations: assessing the nature of responses in word association tasks. *Journal of Psycholinguistic Research* 48.1, 243-256.

Warren, H. C. (1916). Mental association from Plato to Hume. *Psychological Review* 23.3, 208–230.

Warriner, A. B., V. Kuperman, & M. Brysbaert (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods* 45-4, 1191–1207.

Watts, D. J. & S. H. Strogatz (1998). Collective dynamics of "small-world" networks. *Nature* 393.6684, 440–442.

Wettler, M., R. Rapp, & P. Sedlmeier (2005). Free word associations correspond to contiguities between words in texts. *Journal of Quantitative Linguistics* 12.2–3, 111–122.

Wilks, C. (2009). Tangled webs...: Complications in the exploration of L2 lexical networks. In T. Fitzpatrick & A. Barfield (eds.). *Lexical processing in second language learners: studies in honour of Paul Meara*. Bristol: Multilingual Matters, 25–37.

Wilks, C. & P. Meara (2002). Untangling word webs: graph theory and the notion of density in second language word association networks. *Second Language Research* 18.4, 303–324.

Wilks, C. & P. Meara (2007). Implementing graph theory approaches to the exploration of

density and structure in L1 and L2 word association networks. In H. Daller, J. Milton, & J. Treffers-Daller (eds.). *Modelling and Assessing Vocabulary Knowledge*. Cambridge: Cambridge University Press, 167–181.

Wilks, C., P. Meara, & B. Wolter (2005). A further note on simulating word association behaviour in a second language. *Second Language Research* 21.4, 359–372.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review* 9.4, 625–636.

Wolter, B. (2001). Comparing the L1 and L2 mental lexicon: A depth of individual word knowledge model. *Studies in Second Language Acquisition* 23.1, 41–69.

Wolter, B. (2002). Assessing proficiency through word associations: Is there still hope? *System* 30, 315–329.

Woodrow, H. & F. Lowell (1916). Children's association frequency tables. *The Psychological Monographs* 22.5.

Woodworth, R. S. (1938). *Experimental psychology*. New York: Henry Holt & Co.

Wray, A. (2002). *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

Yap, M. J., S. E. Tan, P. M. Pexman, & I. S. Hargreaves (2011). Is more always better? Effects of semantic richness on lexical decision, speeded pronunciation, and semantic classification. *Psychonomic Bulletin & Review* 18.4, 742–750.

Yokokawa, H., S. Yabuuchi, S. Kadota, Y. Nakanishi, & T. Noro (2002). Lexical Evidence, networks in L2 mental lexicon: Japanese, from a word-association task for learners. *Language Education and Technology* 39, 21–39.

Yu, X., Z. Xu, & L. Sun (2011). On Chinese EFL learners' homonym processing in relation to their organization of L2 mental lexicon. *International Journal of English Linguistics* 1.2, 40–49.

Zareva, A. (2005). Models of lexical knowledge assessment of second language learners of English at higher levels of language proficiency. *System* 33.4, 547–562.

Zareva, A. (2007). Structure of the second language mental lexicon: how does it compare to native speakers' lexical organization? *Second Language Research* 23.2, 123–153.

Zareva, A. (2010). Multicompetence and L2 users' associative links: being unlike nativelike. *International Journal of Applied Linguistics* 20.1, 2–22.

Zareva, A. (2011). Effects of lexical class and word frequency on the L1 and L2 English-based lexical connections. *The Journal of Language Teaching and Learning* 1.2, 1-17.

Zareva, A. (2012). Partial word knowledge: frontier words in the L2 mental lexicon. *IRAL: International Review of Applied Linguistics in Language Teaching* 50.4, 277–301.

Zareva, A., P. J. Schwanenflugel, & Y. Nikolova (2005). Relationship between lexical competence and language proficiency: Variable sensitivity. *Studies in Second Language Acquisition* 27.4, 567–95.

Zareva, A. & B. Wolter (2012). The "promise" of three methods of word association analysis to L2 lexical research. *Second Language Research* 28.1, 41–67.

Zhang, D. & Koda, K. (2017). Assessing L2 vocabulary depth with word associates format tests: issues, findings, and suggestions. *Asian-Pacific Journal of Second and Foreign Language Education*, *2*.1, 1-30.

---

[i] Scopus database search and 'documents by year'analysis, 29/10/19, search terms (TITLE-ABS-KEY) "word association" AND linguistic* OR language.

[ii] In this paper we use the term 'L1 speaker' to refer to someone who has acquired the language in question in early childhood and used it (not necessarily exclusively) from early childhood. Please note that the literature we refer to might use other terms such as 'native speaker' or 'L1 user'.

[iii] In word association research, ‘type’ refers to the responses given to a cue by at least one respondent, while ‘token’ refers to the sum total of all responses (or responses of a given ‘type’) to a cue. For example, a cue such as cat might yield 3 response types (e.g. dog, miaow, purr); dog generally accounts for the largest number of tokens.

[iv] In- and out-degree do not sum to the undirected degree because links travel in both directions in a directed network, but are only counted once in an undirected network. In the example given, 4 words (*reading*, *English*, *printing*, and *drawing*) shared both incoming and outgoing links with writing.

[v] The word ‘hub’ is used with two specific senses in this paper. Here, it refers to words which are central to lexical networks (i.e. which have a high in-degree). Later it is used with reference to ‘hub and spoke’ models of semantic knowledge.